Improving Predictability of Cell Culture Processes During Biologics Manufacturing Scale-Up through Hybrid Modeling

by
Zoë Wolszon
B.S., Applied Sciences: Biomedical Engineering, University of North Carolina at Chapel Hill, 2014

Submitted to the MIT Sloan School of Management and the Department of Electrical Engineering & Computer Science in partial fulfillment of the requirements for the degree of

Master of Business Administration
Master of Science in Electrical Engineering & Computer Science

in conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology
May 2020

Signature of Author: _____

MIT Sloan School of Management and Department of Electrical Engineering & Computer Science
May 8, 2020

Certified by: _____
Duane Boning, Thesis Supervisor
Clarence J. LeBel Professor, Electrical Engineering and Computer Science

Certified by: _____
Roy Welsch, Thesis Supervisor
Eastman Kodak Leaders for Global Operations Professor, Management

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

Accepted by: _____
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Committee on Graduate Students

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

Improving Predictability of Cell Culture Processes During Biologics Manufacturing Scale-Up through Hybrid Modeling

By

Zoë Wolszon

Submitted to the MIT Sloan School of Management and the Department of Electrical Engineering & Computer Science on May 8, 2020 in partial fulfilment of the requirements for the degrees of Master of Business Administration and Master of Science in Electrical Engineering and Computer Science

ABSTRACT

In the biotechnology industry, commercial manufacturing of biologic drugs occurs in large-scale production bioreactors (15,000L), but process development occurs in lab-scale production bioreactors (2-3L). Cell culture processes are complicated and the scale-up from bench-scale to commercial-scale can be unpredictable. This study develops an algorithmic approach to better predict the performance of a production bioreactor at commercial scale. A hybrid modeling approach is explored using historical process data and calculated equipment engineering features that characterize the bioreactors at each scale. The study reveals that current process characterization regression models cannot predict commercial-scale performance better than the mean, and that machine learning approaches can improve this performance. Engineering features are found to have a relatively small impact that varies by response variable, but paradoxically are often retained in feature selection of top-performing models. Several new hypotheses arise from these findings, revealing the need for further work with an expanded multi-process multi-scale data set. The researchers propose that by training the model on such a robust data set, it will be possible to test these new hypotheses and unlock significant potential to reduce risk, costs, time, and resources required to develop, commercialize, and manufacture new biological drugs.

Thesis Supervisor: Duane Boning
Title: Clarence J. LeBel Professor, Electrical Engineering and Computer Science

Thesis Supervisor: Roy Welsch
Title: Eastman Kodak Leaders for Global Operations Professor, Management

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

Acknowledgements

There are many people without whom this work would not have been possible. I would like to begin by acknowledging the staff of the MIT LGO program; thank you for allowing me to become part of an incredible community of leaders, for creating constant opportunities to learn and grow, and for pushing all of us to make the most of our time and community. I also want to thank my MIT advisors, Duane Boning and Roy Welsch, who dedicated significant time and effort to helping me shape this project, think through gnarly problems, and refine this long document.

When I began this project, I did not know anything about biomanufacturing, and likely couldn't have told you a single thing about how a bioreactor worked. The team at Amgen welcomed me with open arms, patiently taught me about their world, and took time out of their days anytime I asked to help me with my project. Thank you to my manager, Tom Mistretta, and sponsor, Arun Tholudur, who helped direct the work while being willing to dive into the details with me, and who showed me great compassion. Thank you to my team – Eric Kwei, Kathleen Rand, Maria Perry, Elena Banegas, Michelle Park, Rahul Nechlani, and Larry Sun – who allowed me to pester them with questions, trained me on various systems, helped me find and clean data, and helped me feel at home. Thank you to Aine Hanly for continuously championing the LGO program and being a wonderful leader for AMA and beyond, and to every person who shared their time and brilliance with me – I could not have done this without you.

I was extraordinarily lucky to be able to complete this work alongside several fellow LGO interns; thank you for your friendship and for sharing this journey with me. I also had the great fortune of meeting and learning from an amazing community of LGO alumni at Amgen, all of whom were exceedingly kind, gracious, and excited to share their time and advice – I cannot thank you all enough. Of course, none of this would be possible without Dollie Grajczak, who worked tirelessly to ensure that I was set up for success in every way, and brought together the LGO community at every opportunity. Dollie, we are all so grateful to you for everything that you do.

Finally, I want to thank the other people in my life who inspire me, comfort me, challenge me, and make me laugh on a daily basis: my LGO and MIT Sloan classmates, my friends in Boston and beyond, and my amazing family, without whose love and support I would never be at MIT. In this time of global crisis, you all help me stay connected and make me stronger. I love you all.

The author wishes to acknowledge the MIT Leaders for Global Operations Program for its support of this work.

[THIS PAGE IS INTENTIONALLY LEFT BLANK]

Table of Contents

List of Figures

Chapter 1: Introduction

This study seeks to develop and test a novel way to perform scale-up analyses in biologics manufacturing. This chapter focuses on introducing the biotechnology industry and relevant challenges of biologics manufacturing, presenting the problem statement and hypothesis, and laying out the high-level research methodology. The chapter concludes with an overview of the organization of the thesis by chapter.

1.1 Project Drivers and Motivation

The biotechnology industry – which is responsible for the discovery, development, manufacturing, and distribution of biologic drugs (drugs for which the therapeutic material is produced by living cells) – is currently experiencing a renaissance of growth. Driver include the advent of new technologies from gene editing (such as CRISPR), new potential host cell lines, improvements in biotechnology manufacturing methods such as continuous manufacturing, automation, and more unlock the possibilities of new drug modalities and approaches to treating some of the most elusive and difficult-to-drug targets. Biotechnology companies thus find themselves with enormous opportunities to push the frontiers of drug development and capture enormous value in so doing. However, this opportunity carries with it a cost, in the form of rapidly increasing complexity throughout the discovery, research, and manufacturing stages of drug development.

Developing a new drug is difficult, time-consuming, and incredibly expensive. It is also inherently risky – of those that reach a Phase I clinical trial, only 9.6% of new drugs are eventually approved for marketing and distribution. Simultaneously, as drug pricing continues to be in the political spotlight – with the US House Oversight Committee investigating prescription drug prices and discussions of price caps and potential federal policies on list price disclosures and Medicare reimbursement changes ongoing – biopharmaceutical companies find themselves facing increasing pressure on their margins.[1] The advent of biosimilars compounds pricing pressure, as high-margin therapeutic drugs begin to face competition. However, biosimilars don't just pose challenges for the original drug manufacturers; for the companies developing them, the need to

reverse-engineer the product and maintain tighter specification ranges than normal creates additional constraints on the manufacturing processes, which must meet higher standards of predictability and consistency.

As a leading biotechnology company with a robust, multi-modality pipeline that covers multiple therapeutic areas and includes new biosimilars, Amgen Inc. is strongly affected by these factors. In the face of this increasing complexity, increasing competition, the need for faster time-to-market, and downward cost pressure, operational excellence becomes all the more critical, as does reducing areas of risk in the drug development process. It is in this context that the motivation for this project arises: to improve predictability across scales of manufacturing. The ability to predict commercial-scale performance from lab-scale data and equipment information would provide significant value in mitigating risk, decreasing the time and expense required for repeated not-for-human-use manufacturing runs (for process validation and adjustment), and freeing up commercial manufacturing capacity in a constrained network.

1.2 PROBLEM STATEMENT

The development process for new therapeutics involves manufacturing across different scales of production, and often in different sites and pieces of equipment. For example, commercial manufacturing of biologic drugs occurs in large-scale production bioreactors (15,000L-20,000L), but process development occurs at lab-scale production bioreactors (2-3L) due to the ability to screen and optimize the various process parameters that can influence production in a higher throughput manner. Given that cell culture processes are very complicated, these points of transition from bench-scale to commercial-scale (also known as "scale-up") can be highly unpredictable. This point in the process is also a key source of risk in new drug programs, given that manufacturing sites must be certified by regulatory agencies before they are able to manufacture and distribute a drug for human use, and complications in initial commercial runs can be extremely costly. Even after a drug is being manufactured at commercial scale, business needs often dictate that drugs be manufactured at a different site or multiple sites; in order to properly manage network manufacturing capacity, the flexibility to move these manufacturing processes between sites, scales, and equipment with predictable, consistent performance is critical.

The objective of this research is to explore one potential opportunity to decrease risk, improve predictability, and improve consistency and flexibility in biologic drug manufacturing across Amgen Inc.'s manufacturing network, through a novel hybrid modeling approach to predicting process performance across different scales and equipment.

In support of this objective, this thesis seeks to address the following questions:

- Can advanced modeling techniques *(*such as optimized regularized regression, ensembles of trees, neural networks, and other machine learning techniques*)* improve predictability of select biologics manufacturing processes across scales and equipment, as compared to the application of current linear regression models?
- Does a hybrid modeling approach in which historical process data is supplemented by calculated equipment engineering features improve predictive performance over purely data-driven models when both utilize the same algorithmic techniques?
- Which factors appear to be the most important in predictive models, as determined by feature selection analyses?

1.3 PROJECT SCOPE

This project focuses entirely on a single unit operation and will utilize a single biologic drug as a case study across two scales of interest in order to explore the stated research questions. Specifically, the focus is on the production bioreactor operation at bench scale (2-3L) and commercial scale (15,000L). These scales are chosen for the large differential and their relevance to the high-risk process development step of scaling up to initial commercial production. The unit operation and case study are carefully selected to maximize learnings, address strategically-important challenges, and utilize the maximum amount of process data for a single product. Additionally, focusing on a single product allows for exploration of trends in the process data while controlling for a variety of other factors that would undoubtedly vary widely across products, timelines, and modalities. Given the narrow window of time available, gathering sufficient data to normalize or characterize these other compounding factors is infeasible, leading to the selected approach.

1.3.1 Unit Operation Selection

As stated above, the focus of this research is solely on the production bioreactor. This unit operation is selected for its complexity and critical importance in the biologics manufacturing process. In these processes, the drug is produced by living cells via complex biological processes that are highly sensitive to the physical environment. The production bioreactor is the "heart" of this manufacturing process, where the cells produce the raw material that become the final drug substance in an environment controlled for pH, temperature, dissolved oxygen, and other physiological conditions required by living cells. Thus, the level of productivity and product quality in the production bioreactor is of the utmost importance for all of the downstream processes that follow and the manufacturing process overall. The production bioreactor is also the most difficult to model or fully characterize, and is full of non-linear, complex, dynamic relationships that make scaling the operation up or down a challenge.

1.3.2 Case Study Selection

The case study for this project is selected on the basis of two criteria: wealth of data and manufacturing platform. The product is a single product in Amgen Inc.'s portfolio for which multiple process characterization steps have been completed, offering the best balance of data across the scales of interest, and largest quantity of process data from which to learn. It is also manufactured on a strategically-important platform for Amgen Inc., both historically and moving forward, enabling the project to deliver valuable and actionable learnings to the company.

1.4 Statement of Hypothesis and Research Methodology

As previously discussed, the production bioreactor operation is complex and difficult to fully characterize, in particular due to the time, process, and environment-dependent interaction of physical, chemical, and biological processes. Although the underlying biological processes retain uncertainty and can vary by modality, the engineering features of bioreactors have been studied and well-characterized over time. It is known that while imperfect, equations that approximate mixing, heat transfer, mass transfer, power per unit volume, and mechanical forces within the bioreactor can help scientists and engineers model the macroscopic bioreactor environment.

The primary hypothesis of this work is that adding this type of information to a model that predicts across different scales and equipment would impart a better understanding of scale differences and thus improve predictive performance over the purely data-driven linear regression models used today. The secondary hypothesis is that even absent this additional information, more complex analytical and modeling techniques can better characterize scaling relationships and improve commercial-scale performance predictions over the application of the small-scale linear regression models created during process characterization.

1.4.1 Research Methodology

To evaluate these hypotheses, a subset of both lab-scale and commercial-scale process data from an existing Amgen product is used in concert with calculated equipment engineering features of production bioreactors in the Amgen network. This combined dataset is used to train various algorithms to predict a selection of performance indicators (PIs) – which include measures of process consistency and product quality – at the scale indicated by an input vector of process parameters (which represent the operating characteristics of the bioreactors).

The performance of the models is compared to the data-driven linear regression models currently used for process characterization, which are created and evaluated at small-scale only, as well as compared to a new baseline utilizing various algorithmic approaches without engineering features to evaluate the impacts of both feature addition and different modeling techniques. Features most commonly selected for their predictive power are also analyzed to inform future work on scale-up, and generalizability between two similar processes will be evaluated. A roadmap for future development of the algorithmic tool for scale-up is developed, along with recommendations for improvements in knowledge and data management.

1.5 ORGANIZATION OF THESIS

This thesis considers the application of hybrid modeling for the purposes of improving predictability of production bioreactor performance across manufacturing scales for Amgen Inc.'s Drug Substance Process Development organization. To convey the objectives highlighted above, the document is divided into six sections, in addition to the bibliography and appendices. **Chapter**

1 introduces the project motivation, scope, hypotheses being investigated, and describes the research methodology and high-level approach to the problem.

Chapter 2 provides a brief overview of the biotechnology industry, drug substance manufacturing, process development, scale-up, and technology transfer in the development and manufacturing of a new therapeutic drug. Further discussion is included on the specific role and characterization of the production bioreactor within the manufacturing process. Finally, Chapter 2 provides a view on the landscape of current methods utilized for scale-up analyses throughout the biotechnology industry.

Chapter 3 focuses on the problem formulation and project approach. This includes a discussion of strategic choices regarding project structure and variable selection, an overview of the relevant analytical methods, and an introduction to the machine learning model architecture.

Chapter 4 details the research methodology and in-depth structure and function of the machine learning models. The approach for designing, developing, and testing the models is reviewed, along with details on creating scenarios to evaluate and test each hypothesis. Finally, Chapter 4 discusses key challenges that are encountered and pursuant adjustments made in data pre-processing.

Chapter 5 reviews analysis results, focusing on the relative performance of various models and scenarios in the context of the research questions posed, and evaluating the validity of the hypotheses stated in Chapter 1. Chapter 5 also considers the challenges and limitations of this current research effort.

Lastly, **Chapter 6** summarizes the findings and recommendations arising from this work. Discussions of new hypotheses for future testing and recommendations for future work are included, as are additional recommendations on data-related business processes based on this research experience.

Chapter 2: Background and Literature Review

This study focuses on the process of developing and scaling manufacturing processes for new biologic drugs. This chapter begins with a brief introduction to the biotechnology industry – its history, focus area, and recent trends driving changes in the industry. We also introduce the host company for this research, Amgen Inc. The chapter continues with a detailed overview of biomanufacturing process development and scale-up, as well as an introduction to the production bioreactor. Finally, we discuss the analytical tools commonly used for scale-up operations today.

2.1 A Brief Introduction to The Biotechnology Industry

We begin with an introduction to the biotechnology industry and the host company for this research, Amgen Inc. This section covers the definition and scope of biotechnology, implications on manufacturing of biotechnology products, and key industry trends and their implications.

2.1.1 Industry Overview

Biotechnology is defined by the Oxford Lexico dictionary as "the exploitation of biological processes for industrial and other purposes, especially the genetic manipulation of microorganisms for the production of antibiotics, hormones, etc."[2] Today, biotechnology is a major force in several industries, including agriculture, energy, alcoholic beverages, and more. However, for the purposes of this thesis the focus will be on medical biotechnology.

The medical biotechnology industry began with key enabling discoveries in the early 1950s, including the 3-D structure of DNA and the enzyme DNA polymerase, the establishment of the first continuous cell line (HeLa cells), the discovery of the role of a single amino acid change in causing sickle-cell anemia, and the first artificial synthesis of DNA in a test tube.[3] Modern biotechnology is often considered to have begun in the 1980s, with the first monoclonal antibodies, first vaccine, and focus on the production of proteins as therapeutics.[4]

Since then, the industry has grown in importance and visibility, enabling the treatment of previously incurable diseases and accelerating progress towards personalized medicine. With this has come significant growth in the business of biotechnology: the industry has grown to $112.4

billion USD in the United States alone, and continues to grow at a rapid pace.[5] It is expected to reach $795.7 billion USD worldwide by 2026.[6]

2.1.2 Industry Focus

The biotechnology industry is focused on the discovery, design, development, manufacturing, sales, and distribution of a specific type of drug. Whereas the pharmaceutical industry is focused on small-molecule drugs, which are characterized by a molecular weight of <1,000 Da, often delivered in pill form, and are also known as "synthetics," biotechnology companies create large-molecule drugs, also known as "biologics" or "biologic drugs." These therapies are based on proteins that mimic those created naturally by the body, and must be produced by living organisms in order to replicate the necessary complexity, structure, and biological and chemical attributes. These types of drugs include vaccines, recombinant proteins, biological products such as blood, blood components, cells, and more, and are generally delivered via infusion as opposed to in pill form. They "often represent the cutting-edge of biomedical research" per the U.S. Food and Drug Administration (FDA),[7] enabling treatments for life-altering and life-threatening diseases – often for which no alternatives exist – including advanced autoimmune diseases, various cancers, rare blood disorders, multiple sclerosis, diabetes, HIV/AIDS, and several rare diseases.[8] Due to the ability to target and treat new indications, often with fewer side-effects than broad-acting small-molecule drugs, biologics represent the fastest-growing class of therapeutics on the market and tend to command much higher prices, with the average daily cost of a biologic coming in at >20x that of a small-molecule drug ($45 and $2, respectively) in 2013.[9]

2.1.3 Impact of Molecule Size in Drug Development and Manufacturing

Manufacturing processes for small-molecule and large-molecule drugs differ significantly. While small-molecule drugs are made via chemical synthesis in carefully-prescribed, fully-characterized chemical manufacturing processes, and are defined by a specific chemical formula that can be tested and verified at the end of the process, large-molecule drugs must be made by living cells after the insertion of specific DNA sequences coding the proteins of interest. This requires manufacturers to deal with significant inherent variability and uncertainty that comes with

working with living organisms, and to study and understand the inherent biological and chemical processes surrounding cell growth, protein production, and the factors that affect final drug substance. Additionally, large-molecule drug manufacturing requires very different manufacturing processes, which are focused on growing and nurturing cells, keeping cells in a desired state over a period of time while encouraging them to produce the protein or other molecule of interest, and then carefully purifying the drug substance to remove unnecessary cellular structures and other waste, and to ensure there are no active viral components that could pose a threat to the patient. Because they do not follow a specific chemical formula, these drugs also cannot be characterized by product testing at the end; instead, they are defined and characterized by the exact process by which they are made. Put another way, as is commonly done, "the process *is* the product."

For this reason, characterizing the process carefully to understand the exact controls, timing, and processes that are necessary to produce a drug substance with the intended therapeutic attributes, and then controlling for each of those parameters in the manufacturing of that drug, is critical to producing efficacious, safe, reliable medicines for patients.

2.1.4 Trends in Biotechnology Product Development

As briefly mentioned in Chapter 1, the biotechnology industry is currently experiencing both an explosion of growth and new opportunities and significant challenges that are poised to change the structure and function of the industry. Although these include external factors such as changing regulation with the potential to play major roles in re-shaping the industry, within the context of product development and manufacturing, there are two key trends to understand: the advent of biosimilars, and the exploration of new modalities.

2.1.4.1 Biosimilars

Biosimilars are biologic drugs that are intended to replicate the therapeutic effects of a branded biologic drug already on the market in order to compete with the original branded drug. They are the equivalent of "generic" drugs which come onto the market to compete with branded small-molecule medicines once they go off-patent, and are often associated with significant drops in price. In order to market a generic version of an off-patent small-molecule drug, a pharmaceutical manufacturer only needs to prove that it can reliably produce a drug product that

replicates the chemical formula and properties of the original. This limits the period of exclusivity enjoyed by manufacturers of small-molecule drugs, often viewed as the period during which the company can re-coup the costs of research and development for that drug and all those that failed before making it to market.

For a long period, manufacturers of biologic drugs continued to enjoy exclusivity even after their patent protection expired. Because there was no validated and approved method to prove equivalency of the competitor and the original. As a result, competitor drugs were prevented from reaching the market without undergoing a time-consuming and expensive new clinical testing and regulatory approval process. However, this has changed over the last decade as regulatory agencies have specified processes by which sufficient therapeutic equivalency can be proven. The European Union has led in this regard, with the European Medicines Agency (EMA) publishing its official guidelines in 2005 and granting approval of the first biosimilar in 2006; since then, the EMA has approved more than 50 new biosimilar products. The U.S. FDA has lagged its European counterpart; although it published its official regulatory pathway for biosimilars in the Biologics Price Competition and Innovation Act (BPCIA) passed as part of the Affordable Care Act (ACA) in 2009, the first biosimilar was not approved until 2015, and fewer than 20 new biosimilars have been approved since, making the category a nascent but highly-anticipated part of the US pharmaceutical market.[9,10] It is worth noting that biosimilars still do not currently pose the same type of competitive threat to original drug manufacturers as generics, primarily because the two products are not *exactly* the same (hence the name bio*similars*). The inherent complexity and production process of the drug substance can impart different biological properties and thus different side effect profiles to the different biologics, which, along with unique structural properties of the U.S. healthcare market, often results in higher stickiness of the original product for patients already taking the branded therapeutic and hesitant adoption from some clinicians and patient populations. This is in addition to the protection afforded from the manufacturing complexity and uncertainty.

That said, biosimilars are undoubtedly having their impact on the biotechnology industry, and their effects are only expected to amplify in the coming years as biosimilars become a bigger part of the US therapeutic landscape. The advent of biosimilars has two major impacts on the

industry. First is increased competitive pressure – for manufacturers of biologics that have expired patent protection or are soon going off-patent, this introduces competitive pressure in a way that did not previously exist, along with a race to create competitors to blockbuster biologics across therapeutic areas. Second is the increased importance of tight controls in biologics manufacturing, both for companies trying to bring biosimilars to market and those working to develop new branded biologics with the maximal level of protection.

For companies working to create and bring a biosimilar to market, the development process is not as straightforward as one might think. Unlike small-molecule drugs, there is no "formula." Since the product is the process and the details of the original drug's manufacturing process are considered trade secrets and thus kept confidential, companies have to reverse-engineer the product, using their knowledge of the therapeutic protein and significant internal process testing to determine how a product with the same therapeutic effect can be produced. Additionally, the manufacturing specifications – which are set by the company – must be tighter than normal in order to replicate the original drug's process as closely as possible. In fact, these process parameter and performance indicator specification ranges are a key part of regulatory approval for biosimilars. Setting these ranges appropriately requires educated guesses of competitor processes along with a detailed understanding of the capabilities of internal manufacturing processes to manage risks and costs associated with deviations, non-conformances and other irregularities which require investigations, process adjustments, and regulatory reporting. This is also a reason that excellence in biologics manufacturing can create a significant advantage for companies not only in developing and launching biosimilars, but also in the development of new biologics that will benefit from the protections of having very tight specifications for the original product that competitors will have a hard time matching.

2.1.4.2 New modalities

Another major trend in biotechnology drug development is the increasing number of new modalities being explored by start-ups and large companies alike. Modalities are unique types of therapeutic drugs characterized by the type of protein being produced and its mechanism of action in the body to achieve the desired therapeutic effect. Examples of modalities include monoclonal

antibodies (mAbs), bi-specific T-cell engager (BiTE®) molecules, Car T Cells, fusion proteins, peptibodies, oncolytic immunotherapy viruses, peptides, and more.[11]

In addition to being an exciting area of opportunity for biotechnology companies to find solutions to diseases that have to-date proven impossible to cure or manage, the proliferation of new modalities in a company's research and development pipeline and product portfolio introduces significantly more complexity in managing manufacturing technologies, platforms, and networks, along with the increase in uncertainty that accompanies attempting to manufacture a new-modality drug for the first time. In this context, the ability to predict manufacturing performance and flexibility in the manufacturing network – and thus the ability to move production of a particular biologic from one site or area to another, and to swap in new technologies, controls, raw materials, and more – is critical.

2.1.5 Introduction to Amgen Inc.

Amgen Inc. ("Amgen") is the world's largest independent biotechnology company, and a leader in biology-first, modality-independent drug discovery, development, and manufacturing. Amgen's mission is "to unlock the potential of biology for patients suffering from serious illnesses by discovering, developing, manufacturing and delivering innovative human therapeutics," and the company focuses specifically on areas of high unmet medical need. Since its inception in 1980, Amgen has grown into a company with a market capitalization over $100 billion and $23 billion in revenue in 2019. Amgen currently has 23 products on the market spanning six therapeutic areas and presence in approximately 100 countries around the world.

With an oft-quoted core value of "every patient, every time," Amgen places operational and manufacturing excellence at the same level of importance as scientific innovation, and sees biomanufacturing expertise as a key differentiator and primary endpoint of everything they do.

Amgen is frequently recognized as having one of the most robust pipelines in the industry, with 45 drugs currently in clinical trials spanning various modalities, therapeutic areas, and indications. With a clear strategic focus on "World-Class Biomanufacturing" (listed as one of Amgen's three strategic pillars) and several manufacturing sites across the globe leveraging

multiple manufacturing platforms, speed, flexibility, and capacity in the manufacturing network are critical to Amgen's continued growth and success.

2.2 PROBLEM BACKGROUND: PROCESS DEVELOPMENT AND SCALE-UP

This project is focused on the manufacturing of biologic drugs, and more specifically, the creation and execution of carefully-controlled manufacturing processes across scales of manufacturing necessary to safely and reliably produce the product. As such, it is necessary to provide key background information on these products, and what it takes to manufacture them successfully.

2.2.1 Summary of Biologics Drug Substance Manufacturing

The process of manufacturing a final biologic drug product for sale requires manufacturing or acquiring several items, including the drug substance, drug product, and the drug delivery technology. The drug substance is the active therapeutic substance, generally in liquid form, that will be applied, consumed, or injected into the human body. The drug product is the final form factor that will be delivered – this involves putting the drug substance into the final form (most commonly a liquid for infusion in the context of biologics), container, packaging, labeling, and drug delivery technology with appropriate instructions. The drug delivery technology itself could be a glass vial to be used with a syringe, a pre-filled syringe, a commercially-available delivery device, or a proprietary drug delivery technology that is paired with the drug itself. The focus of this thesis will be solely on the drug substance manufacturing.

Drug substance manufacturing is divided into upstream and downstream processes. The high-level definition of the two processes is as follows. Upstream processes are focused on the production of the desired active drug substance by living cells. This involves the preparation of raw materials, media, and feeds and all cell culture processes, including the seed train expansion stage during which the cells are encouraged to grow and moved into subsequently larger vessels until they reach the production bioreactor. The production bioreactor unit operation is the "heart" of the biomanufacturing process, as it is focused on the production of the raw drug substance. Downstream processes are generally focused on purification of the drug substance, and are often

sub-divided into four steps: preparation, capture, purification, and processing.[12] This includes product extraction and capture, concentration, size and charge-based filtration, viral inactivation, and more. The objective is to arrive at a final drug substance that is ready for human use in all ways except for the final form factor, delivery technology, and associated packaging. The two processes are separated by the harvest step, which is focused on extracting the substance of interest from the production bioreactor at the end of its operation and passing that substance down to the downstream filtration steps. These steps are laid out visually in Figure 1.[13]

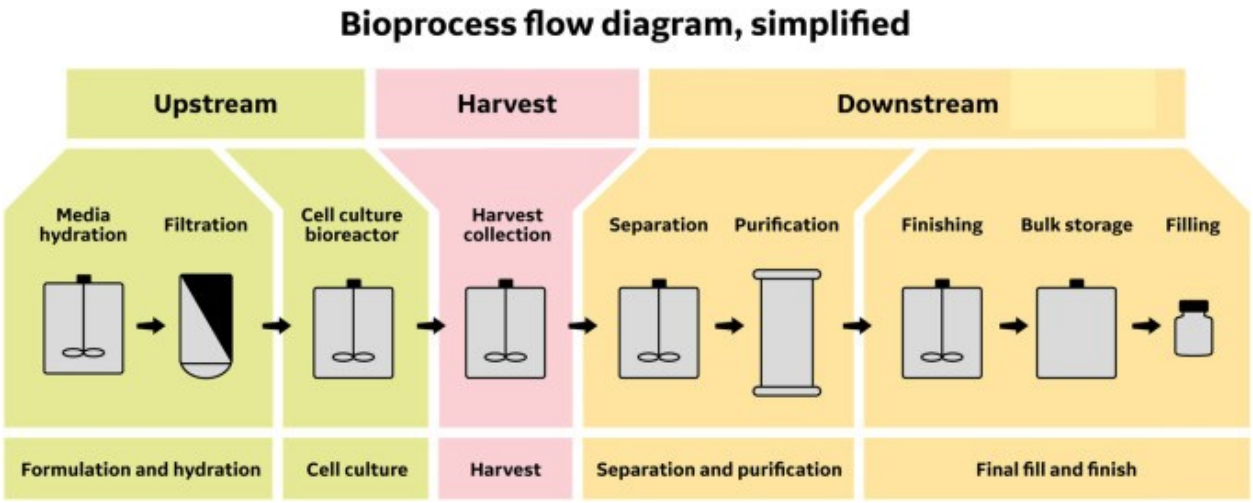

Figure 1. Simplified View of Biomanufacturing Process. The biomanufacturing process is divided into upstream and downstream activities (separated by the harvest step, which collects the raw material at the end of the production bioreactor unit operation). It is worth noting that the Drug Substance portion of manufacturing ends after purification. The "Final fill and finish" category is generally regarded as Drug Product.

2.2.2 Introduction to Process Development

Process development is a critical piece of developing, producing, and generating revenue from a new drug. Chronologically, it falls between research and development activities, where new drugs are discovered, and manufacturing and commercial activities, where drugs are commercially produced, sold, and distributed. It concerns itself primarily with characterizing the exact process by which a drug is produced, which is, as discussed, critical for biologic drugs, where the process essentially *is* the product (as opposed to a chemical formula). It also seeks to create, define, and characterize a manufacturing process that is both effective at achieving the precise product profile and set of attributes desired and that is as efficient, well-suited to the company's facilities and capabilities, and well-controlled as possible.

The process is organized around four primary steps, each with their own objective as laid out below.

1) *Early Development and Clinical Material*: create an initial process to ensure the drug is manufacturable and create material for clinical and toxicological study.

2) *Commercial Process Development (CPD) and Process Characterization (PC)*: explore, understand, optimize, and characterize the manufacturing process.

3) *Technology Transfer*: transfer the finalized manufacturing process from the lab to a commercial manufacturing facility, or from an initial commercial manufacturing facility to a different facility and/or site; test process, validate performance, and characterize process in new site.

4) *Commercial Manufacturing*: support regulatory certification of process in commercial manufacturing facility and begin producing drugs for commercial sale and distribution.

The following sections will go through each of these steps in more detail.

2.2.2.1 Early Development and Clinical Material

In the initial stages of development – specifically, the pre-pivotal phase, before first-in-human (FIH) trials – the optimal process of producing a given product has not yet been set. In this phase, there is one known method for producing the therapeutic molecules with the desired properties based on discovery and research and development efforts, and that method is utilized to produce material for testing (generally at small scale) along with toxicology studies and clinical materials (usually produced at an intermediate or commercial scale) without robust experimentation as to the process. This stage provides valuable information in the form of material for testing, initial exploration of the qualities of the therapeutic molecule and a way to manufacture it successfully, and, most relevant to this project, the initial data from manufacturing at a larger scale (either commercial-scale or an intermediate scale).

2.2.2.2 Commercial Process Development and Process Characterization

In this next stage of process development, which generally occurs after or concurrently with clinical trials (and thus is referred to as “pivotal”), scientists work to replicate the product while

optimizing the process, characterizing the design space within which the production will be successful, and setting limits for process parameters and performance indicators to ensure that the final drug substance will meet quality and safety requirements, along with all other specifications. Commercial Process Development (CPD) is focused on the first task – optimizing, studying, and finalizing the process itself – whereas the ensuing Process Characterization (PC) effort is focused on characterizing the process and setting all relevant limits, target values, and specification ranges. The primary methodology for this process characterization process is a structured approach known as the Design of Experiments (DOE) method.

2.2.2.2.1 Design of Experiments (DOE)

Design of Experiments is a structured approach to characterizing the design space of a manufacturing process by perturbing variables in a structured way in order to discern the relationships and impact of individual process parameters and interactions between process parameters on the final performance indicators. Multiple linear regression analyses are then performed on the results in order to characterize these relationships between inputs (process parameters) and outputs (performance indicators), and based on this, to set operating ranges and specification limits for each parameter.

This method of process characterization is widely-used throughout the industry for a few reasons. In a world of limited time and resources, it is not feasible or desirable to run an experiment for every possible value and permutation of a group of process parameters; DOEs enable scientists to efficiently test variable ranges of parameter values and permutations in a way that lends itself to broader understanding of the process while allowing customization based on experience, intuition, and process knowledge. It is a well-studied and characterized method that, with proper application and analysis, can elucidate not only effects of perturbing a single parameter but also the effects of interaction terms, and can help scientists understand which parameters are critical to control tightly and which do not appear to have a significant impact. Finally, it lends itself to the accumulation of prior knowledge based on the analysis of patterns across DOEs, which can then be used to inform future experimental plans. For all of these reasons, DOEs become a tool of not only understanding, but also significant gains in efficiency with appropriate risk mitigation.

The DOE can be supplemented by one-off "challenge studies" which are intended to test values for a particular parameter or set of parameters that fall outside the normal range of what would be tested in the DOE. These are often used to explore whether it would be possible to expand the range for a given parameter. For the purposes of this research project, data from these challenge studies are used in concert with the DOE experiment data in order to capture additional information about the impact of process parameter variability; these studies are often the sources of the outliers in the model inputs, as will be discussed in Chapter 3.

The product of the Process Characterization (PC) study is a Process Characterization Report, which is intended to share the outcome of the experiments with a particular emphasis on 1) identification of which parameters and interaction terms (where applicable) are "critical", 2) the operating ranges and specification limits of both inputs (process parameters) and outputs (performance indicators) of the production bioreactor to ensure that the final drug substance is within the target product profile (final specifications), and 3) the outputs of multiple linear regression analyses on each priority parameter characterizing its impact on key performance indicators.

However, it is important to understand that these regression analyses are not currently used to directly predict commercial-scale performance. Instead, they are used to characterize variability around the set operating ranges and specifications; this is then used to set ranges around expected commercial setpoints for normal operation, along with action limits, specification limits, and other indicators for operators and process monitoring specialists. It is in the initial technology transfer stage that differences between these expected ranges and the actual ranges are usually identified; these often take the form of an "offset" which is then used to adjust the process parameters, operating ranges, and limits specified during process characterization.

In addition, process characterization reports include regression analyses "with scale" that are of particular relevance to this project. These analyses are performed on not only the small-scale DOE data, but also the initial intermediate- or commercial-scale manufacturing runs used to produce material for toxicology and clinical studies (often limited to only 1-3 runs). The purpose of these analyses is to identify whether there are particular parameters or performance indicators that appear to have significant scale-dependence.

It is notable and of particular relevance to this project that these analyses are *not* currently used to predict performance of the given process at a different scale. This is due to the limitations of the current analytical techniques; in particular, multiple linear regression does not perform regularization or feature selection of inputs. In this case, with a dataset with many variables and a relatively small number of observations, this leads to the assignment of a coefficient for every single variable (many of which are collinear), and is particularly susceptible to differences in magnitude. In many cases, a standard linear regression on the full dataset with engineering features was assigning coefficients of +e25 and -e25 to each of the explanatory variables. Thus, it did not appear to be the best technique for this particular study. Additionally, as was discovered in the course of this project, making cross-scale predictions with only a select few commercial-scale runs from which to learn is very difficult without a robust analytical understanding of scaling relationships.

2.2.2.3 Technology Transfer

Once the processes are fully characterized by Process Characterization, it is time to transfer them to facilities for commercial production. In order to do this, it is necessary to move the processes to different equipment at a larger scale, and often to a different facility and/or site (depending on factors such as availability of a Good Manufacturing Practice, or GMP, facility, capacity, and desired scale of initial commercial production). Even after initial commercial production, a product may be transferred to a different facility or site within a company's manufacturing network. Regardless of the exact point in the lifecycle of the drug, this process is known as technology transfer; the technology being transferred is the manufacturing process for the new biologic drug.

The process for technology transfer may vary based on the product, facility, site, and in particular, the relative scales of the originating and destination facility. After the transfer of knowledge, reports, and specifications, the process will need to be validated in the new facility. This is generally done through the execution of not-for-human-use (NHU) or "engineering" runs. As indicated by the name, these manufacturing runs are used for testing and process validation (and potentially adjustments), and the resulting product is not able to be sold, given as samples, or in any other way purposed for use in a human. These runs are not required by regulatory standards,

but are common practice as a risk management strategy, for training and validation, and to ensure that the facility is ready for production and Process Performance Qualification (PPQ), which will be discussed in the next section.

These engineering runs may be completed in the final commercial-scale equipment and/or in a validated small-scale model in the facility. This small-scale model is often at an intermediate scale, and has been previously validated to be representative of the commercial-scale equipment in that facility, with clearly-defined and understood relationships and scale differences. Based on the process, familiarity, initial results, and assessed level of risk the number of engineering runs required can range from 2 to >10, and can be only in the small-scale model or in both the model and the full-scale production equipment.

It is central to this project that engineering runs, particularly high numbers of them, are very resource-intensive and expensive. They also are non-revenue-generating manufacturing runs that take up valuable capacity in a manufacturing facility, in an industry where a single manufacturing run often takes multiple weeks and produces millions of dollars' worth of product. These engineering runs thus result in a net decrease in the total throughput of a manufacturing facility, potentially increasing total costs of goods manufactured, add to the cycle time for regulatory submission, and can delay product launch or threaten meeting current demand, which is of the utmost importance within Amgen (as one may recall, "every patient, every time" is a core principle). Since engineering runs are essentially a risk management strategy focused on validating that the process will perform as expected in the destination equipment and managing process adjustments as problems arise, it is expected that engineering runs could be greatly reduced with better predictability across scales. Along with reduction of risk at this point in process development, a primary endpoint of the long-term implementation of this project would be reduction of engineering runs and accompanying cost, capacity, and resource savings.

2.2.2.4 Commercial Manufacturing

2.2.2.4.1 Process Performance Qualification (PPQ)

After technology transfer and before full commercial production, it is necessary to officially validate the consistency, reliability, and accuracy of the production process with the relevant

regulatory bodies, in the equipment and facility in which it will be commercially produced. This takes the form of a PPQ, during which the facility produces three production lots that are tested at every unit operation to ensure that they remain within the specification limits and are consistent across lots. If the PPQ is successful, then the facility is approved for commercial production of the specified product, and the three lots produced during the PPQ process are able to be sold and distributed.

However, if the PPQ is not successful – based on differences between the lots, biocontamination, non-conformances, or other issues – the facility will not be approved and will need to launch investigations and process improvements to remedy any identified issues before attempting PPQ again. This requires more engineering runs, significant resources for investigations, process adjustments, additional testing, reduced manufacturing capacity, potential reputational risk with regulatory bodies, a delay in getting a new product to market and/or fulfilling current demand, and more. Additionally, all material produced during the PPQ process is not able to be sold or distributed for human use. Thus, failing PPQ is very costly and significant time and effort goes into ensuring this does not happen.

As a result, technology transfer, scale-up, and PPQ represent critical points of high risk in the development, manufacturing, and lifecycle management of a biologic drug. All hinge on detailed process understanding and performance predictability within and across specific scales, facilities, and pieces of equipment.

2.2.2.4.2 Commercial Production

In this stage, the manufacturing process has been fully specified and the performance validated, so the facility is in pure production mode. Throughout commercial manufacturing, the process is continually monitored to ensure that process parameters and performance indicators stay within the specified ranges throughout the process for each batch. Any deviations, non-conformances, and trends are identified, discussed, and appropriately investigated. If need be, expected ranges for a given performance indicator may be adjusted through the proper facility, company, and regulatory channels.

2.2.3 Role of Scale-Up in Process Development

It is important to understand the order of magnitude of each scale of production involved in biomanufacturing, as well as the differences between them and resulting key role of scale-up and scale-down operations. For each step in process development detailed in the previous section, there is a typical scale of production associated with that step. For the unit of operation of interest here – the production bioreactor – that scale is generally discussed in liters, to represent the internal volume of the bioreactor during production. The exact scales vary based on the selected biomanufacturing method, platform, and equipment. For example, biologics produced using single-use bioreactors (SUBs), which utilize single-use bags to contain the product and process, or continuous manufacturing processes, which is a single closed system with continuous flow, will generally operate at much smaller commercial production scales than fed-batch processes that utilize stainless-steel bioreactors.

For this project, the process of interest is on a fed-batch platform that utilizes stainless-steel bioreactors. For this manufacturing method and selected equipment, the scales considered range from 2-3L at small-scale, which takes place on a lab bench, up to 15,000-20,000L at commercial manufacturing scale. As a result of the varied scales involved in process development, every new process must be scaled-up, scaled-down, and transferred among various equipment, facilities, and scales repeatedly during its development. This presents additional challenges and points of risk in the process, as every transfer adds uncertainty and requires performance validation in the new scale and equipment. The scales involved in a typical manufacturing process similar to the one investigated in this project are shown in Figure 2, below. Note that all scales are descriptive of the volume of the production bioreactor unit operation.

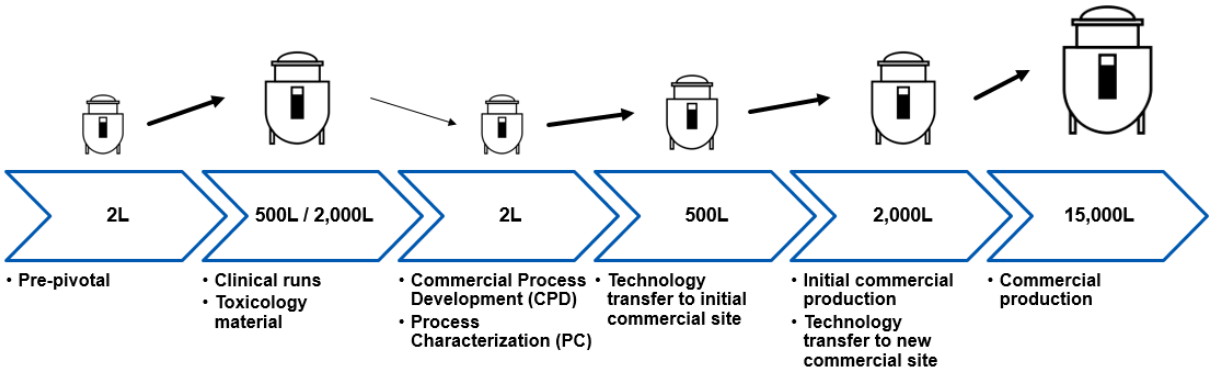

Figure 2. Scales of Production in Biomanufacturing. This figure depicts the scales of production involved in the development of a new therapeutic drug manufacturing process. All scales shown are specific to the production bioreactor unit operation, and shown in liters to represent the volume of the bioreactor. Exact bioreactor volumes vary; numbers shown here representative only.

2.2.3.1 Scale Differences across Unit Operations

Each unit operation scales in a unique way, and thus must be modeled, understood, and validated across scales on its own. Certain unit operations may scale in a more predictable, mechanistic way than others; for example, chromatography columns are fairly well-characterized by physical and chemical equations. However, as will be explored in the next section, the production bioreactor is a complex, heterogeneous operation seriously affected by biological processes that have not been fully characterized, making predictable scale-up a challenging task.

2.2.4 The Production Bioreactor in Biotechnology Manufacturing

In the production of large-molecule drugs, the production bioreactor (also referred to as the N-0 step) is the "heart" of the manufacturing process. This is where all of the raw material that will become the drug substance is created by living cells. These cells are grown from a vial in a series of shake flasks, wave bags, and smaller bioreactors until they are of a sufficient density, and then transferred to the production bioreactor for production of the protein of interest. Once the N-0 step is complete, the rest of the manufacturing process is focused solely on harvesting the protein of interest and purifying it into the final drug substance. Thus, the production bioreactor is the most critical and sensitive step in the process, and for this reason, it is carefully controlled and monitored.

2.2.4.1 Basic Anatomy

A production bioreactor is made up of a number of key parts, each of which have a significant impact on the controls and functioning of the reactor. For the purposes of this thesis, the description will remain at a high-level and will focus on only three components that are necessary to understand the basic functioning of the bioreactor: the vessel, the impeller(s), and the sparger.

The vessel is the container itself, and is generally a cylindrical shape made of stainless steel (even in the case of single-use systems, which this paper will not discuss explicitly). The cylinder has a number of specific characteristics and built-in features, such as a jacket that is responsible for maintaining thermal stability and adjusting temperature with circulation of a fluid, and a number of specialized ports for sampling, harvest, probes, and more. The vessel is pre-seeded with media (on which the cells can feed to obtain the necessary nutrients for growth and protein production) before being inoculated with the cells that will produce the protein of interest. Based on the selected manufacturing platform (batch, fed-batch, or continuous manufacturing), the bioreactor contents may then stay constant, receive boluses of media, feed, glucose, and other nutrients and metabolites throughout the process, and/or be continuously supplemented and drained in small volumes throughout the length of the operation.

The impeller system is a specifically-designed mixing tool intended to circulate the cells and fluid within the bioreactor to achieve as homogenous a mixture (and thus environment for the cells) as possible. This system can be a single impeller or multiple impellers depending on the size and design of the bioreactor, and impellers can rotate on the central axis to move fluid around the reactor within a given layer and/or move up and down to move fluid between the vertical layers of the reactor. Generally, the larger the vessel, the more impellers are required to achieve maximum homogeneity. There are many different impeller designs from which manufacturers can choose – a non-exhaustive list of example impeller types for the curious reader who would like to further investigate include marine, elephant ear, Rushton, pitched-blade, spin filter, and cell-lift impellers – each of which have their own geometries, patterns of motion, and impacts on the fluid dynamics, forces imparted on the cells, and the overall solution. The rate at which the impellers move is generally referred to as the "agitation rate," which is a process parameter set and controlled in manufacturing processes.

The sparger is responsible for inserting oxygen, nitrogen, carbon dioxide, and/or any other gases desired into the solution in order to provide the desired level of dissolved oxygen to fuel cell growth and metabolism and to control the pH of the solution. This is accomplished by pumping the desired gas into a sparger, a specially-designed pipe with a specified pattern of holes in it (or simply an open end), where each opening releases its own stream of bubbles into the liquid. The size and exit velocity of these bubbles is important, since bubbles that are too small and moving too fast can rupture and thus kill cells, but bubbles that are too large will not be available to cells for uptake, and bubbles that are moving too slowly may not reach the top layers of the solution. The type, quantity, and flow of these gases is generally controlled by a PID system with the objective of controlling a set level of dissolved oxygen in the solution (in addition to maintaining the specified pH setpoint, which is accomplished through the sparging of CO_2). See Figure 3 for a simplified diagram of a typical bioreactor.[14]

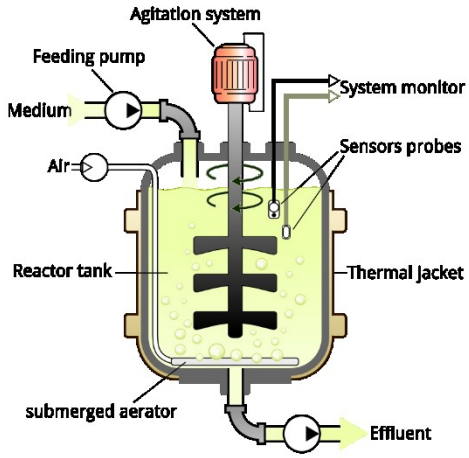

Figure 3. Simplified Bioreactor Diagram. Diagram shows the key components of a typical bioreactor such as the tank, aerator, agitation system (with three impellers shown), feed pumps, probes, and thermal control system.

2.2.4.2 Equations, Approximations, and Characterization

As a result of their importance in the manufacturing processes for biotechnology products, fermented products, and certain chemical products, bioreactors have been extensively studied. Within the context of biotechnology manufacturing, these studies have largely focused on ways to characterize the physical, chemical, and biological properties of the internal environment. This internal environment is inherently complex – it is undoubtedly heterogeneous, affected by a

number of interdependent factors. For example, a small subset of interdependent factors would include volume, impeller power, agitation rate, shear rates created by the rotation of the impeller on cells located next to the impeller, eddy creation and size, sparge volume, exit velocity of sparged gas, gas bubble size, oxygen uptake rate of the cells, and metabolite concentrations that signal information about the status and level of stress of the cells.

2.2.5 Current Methods for Scale-Up Analyses

Scaling up biomanufacturing processes is both an art and a science, and there is not one generally agreed-upon method within the scientific community or the industry by which to scale a process. Regulatory agencies provide some guidance, but no directives as to what method should be utilized.[15] As a result, several approaches to scaling analyses exist, and the most predominant approach for a given process may vary based on company policy, experience, and expertise, the modality, equipment, and scale, and which parameters, phenomena, and/or product quality attributes are the most critical or potentially problematic.

However, several of the key approaches focus on mixing; per Braem et al. in their guide to process scale-up and assessment, "Generally speaking, the mixing power is much greater in the plant than in the laboratory. It can be challenging to simulate the mixing behavior that will be obtained on scale in the laboratory, since there are many variables one can choose to hold constant between the experiment and the plant run. […] Different phenomena scale with different variables, and it is not always well understood which variable is the best choice for scale-up and scale-down."[16]

2.2.5.1 Commonly-Used Analytical Approaches: P/V and Impeller Tip Speed

Two of the most common analytical methods – which maintain a focus on mixing – are maintaining a constant power per unit volume (P/V) and a constant impeller tip speed. Each of these two methods intends to scale by keeping a selected key parameter constant across the two settings, with the idea that maintaining that critical parameter will ensure the most similar environment for the cells in the bioreactor.

In the case of power per unit volume (P/V), this becomes a clear, easy-to-understand, and linear scaling approach. With a set bioreactor volume and a known impeller pumping number (a

property of the part itself), the process development scientist needs only to adjust the agitation rate to the appropriate level to achieve the same power per unit volume. However, this method is not selected due to simplicity alone. P/V is an important driver of the cell culture environment – power incorporates several properties of the impeller in concert with constants and the agitation rate of the process and affects the average shear rate and eddy size, while volume is a clear scaling factor between bioreactors. Additionally, the combination provides an indication of the consistency of the cell environment and is related to the mixing sufficiency (to achieve a homogeneous environment, or as close to one as possible). The P/V formula is shown in Equation 1.

$$\frac{N_p * \rho * {D_i}^5 * N^3}{V}$$

Equation 1. Power per unit volume calculation for bioreactors. N_p = impeller power number (dimensionless), $\boldsymbol{\rho}$ = cell culture fluid density (assumed to be equivalent to water at 1000 kg/m^3, and constant across scales), D_i = diameter of the impeller (meters), N = process agitation speed (revolutions / second), and V = process working volume of the bioreactor (m^3).

The other primary analytical method commonly used in process scaling is maintaining a constant impeller tip speed. This method has several things in common with the P/V method, in that it is focused on a mechanical property that drives mixing, homogeneity throughout the bioreactor, and mechanical forces imparted on the cells. In these ways, it is largely a simplified and reduced version of P/V scaling, since if the bioreactor set-up, impellers, and working volume are constant, the tip speed will capture the same information. Of these factors in a constant manufacturing environment, the working volume is the most likely to change by process; as such, evaluating the impact of both methods may be useful.

2.2.5.2 Other Methods

Other methods for scale-up assessment include 1) geometric ratios, such as the diameter of the impeller to the diameter of the reactor; 2) similar mixing-related mechanical properties, such as shear, flow per volume, torque per volume, Reynolds number at the impeller tip, and blend time; or 3) biochemical properties that affect availability of oxygen and other key chemicals for the cells in the reactor, such as k_La and the characterizations of the microenvironment (focused on pH and relative concentrations/partial pressures of CO_2 and O_2). The k_La metric is particularly powerful in terms of characterizing the environment within the bioreactor for the cells. However, it is

extremely resource-intensive to study at high-fidelity (through computational fluid dynamics modeling, other advanced modeling, or actual experiments in the reactors), is rarely fully-characterized in small-scale reactors, and can vary based on a number of parameters such as tubing used (particularly relevant for small-scale experiments), making it a difficult primary strategy for scaling.

2.2.5.3 Exploration of Multivariate Analysis for Scale-Up

Given the number of potential methods for scaling up a process, and the many ways to characterize different parts of a process, multivariate analysis is gaining interest as a different way to scale-up a process. This would involve leveraging all of the data created in small-scale experiments throughout process development and characterization – inputs, intermediate measurements, and outputs – and using more advanced statistical techniques to discern valuable information about the process and the relationships between the inputs and outputs. Mercier et al. showed value in the exploration of the data sets using Principal Component Analysis (PCA) for clustering purposes, but struggled with insufficient fit and predictive power of the attempted Partial Least Squares regression (PLS). They attributed this primarily to the lack of robustness and structure in the data itself (insufficient variability in key parameters to capture their impact and insufficient parameters to capture all key attributes), particularly in early data sets (prior to the Design of Experiments stage). They argued that multivariate data analysis "should be routinely used to analyze early development data to reveal relevant information for later development and scale-up" in concert with more robust, structured experimental data sets early on in process development.[15]

2.2.5.4 Implications for this Project

It is clear that the biopharmaceutical industry continues to struggle with the question of how to reliably, robustly, and efficiently scale-up a biomanufacturing process in a standard way, and is eager for more advanced analytical approaches to assist in this. In many ways, this project is a continuation of the exploration of multivariate data analysis with early process development data as discussed above. In this project, that data is expanded to focus on cross-scale differences including the final commercial scale of production, with a focus on supplementing the data available with calculated engineering features – in hopes of more effectively characterizing the

key attributes of the process – along with the use of more advanced algorithmic techniques intended to elucidate key relationships between inputs and outputs across scales.

Chapter 3: Problem Formulation

Having now established the context and need for a more robust analytical solution for biomanufacturing scale-up processes, we turn now to the formulation of the problem. This project focuses on leveraging Amgen Inc.'s historical process data, equipment information, and significant expertise in process development and best-in-class biologics manufacturing to gain insights into the factors that drive bioreactor performance across scales and equipment. Accomplishing this requires surveying available data and expertise, selecting the process parameters and engineering features to include, determining which indicators of performance are most valuable to the organization, and creating an appropriate machine learning architecture to enable robust analysis with the given data structure. This chapter walks through this process, with one section dedicated to each step in the stated order.

3.1 Overview of Available Data

The historical process data utilized in this project is taken from measurements of the production bioreactor parameters, environment, volume, and associated analytic assays. These measurements are a required part of the manufacturing process (as governed by internal policies and external regulatory bodies), as they are essential to monitoring, measuring, and controlling all variables which could impact the drug product. The data for this project is from two sources: 1) experiments at small-scale (2L and 3L bioreactors), which entail significant variation in process parameters in order to discern relationships between process parameters and performance indicators; and 2) manufacturing runs at commercial-scale (15,000L bioreactors in this case), where process parameter setpoints do not vary and the controls are maintained within tight specifications, including with closed-loop PID systems for certain parameters.

The data available for commercial-scale processes is robust – there are thousands of unique observations providing data points for every measurement (of process parameters, concentrations, and indicators of status, growth, cell stress, and many more) for each individual batch, at regular time intervals in the multi-day or multi-week-long production bioreactor unit operation. However, this type of standardized, robust, and complete time-series data is much less available for small-

scale experiments, and thus it is difficult to match the two data sets in a comparable way beyond initial measurements, final day measurements, and certain intermediate measurements. Given this, the focus for the project is on individual point measurements (as compared to reduced-dimensionality time-series data), and in particular on initial and final measurements.

3.2 Selection of Response Variables

This project considers a total of three performance indicators as response variables in the analysis. Performance indicators are "outputs" of each unit operation in the manufacturing process, which include measures of both process consistency and product quality. Examples of process consistency and efficiency metrics include Final Titer (a measure of the concentration of the molecule of interest in the final bioreactor volume), Final Viability (a measure of the percentage of living cells), Specific Growth Rates (measures of cell growth over a specified period of time), Final Viable Cell Density (VCD) (a measure of the density of living cells in the final bioreactor volume), final concentrations of metabolites, integrals and derivatives of each of these measures, and more. Examples of product quality attributes include glycosylation metrics (measures of glycoform profiles which can have significant impacts on efficacy of a protein-based therapeutic), analytical assay results such as CEX-HPLC Acidic, Main, and Basic Peaks (which test the charge density), SE-HPLC High Molecular Weight (which interrogates the molecular weight distribution to identify and flag potential issues with aggregates), and many more.

Each of the three performance indicators selected is chosen from the full set of potential response variables based on the criteria of process importance (as found in process characterization reports of the processes of interest), scale dependence, reliability of measurement, and criticality and level of interest from process development scientists at Amgen Inc.

3.3 Selection of Explanatory Variables

Process parameters are the controls, or operating characteristics, of a biotechnology manufacturing process. For the unit operation of interest here, the production bioreactor, these are the parameters that intend to control the internal environment in an exact, pre-specified way so as

to encourage optimal performance, growth, and product quality (the factors measured by the performance indicators discussed in the section above). These include parameters such as temperature, pH, concentration of various metabolites, dissolved oxygen, agitation rate (the rate of rotation of impellers designed to mix the contents of the bioreactor in order to achieve a homogenous solution), and several more.

For the purposes of controlling the size and shape of the data set for analysis (and ensuring a reasonable ratio of observations to features), a small subset of seven process parameters is selected to make up the set of explanatory variables for these analyses. These parameters are chosen based on demonstrated importance and level of interest, and are selected on the basis of consistent data availability, interviews with Amgen scientists, review of previous process characterization studies, and literature review.

3.3.1 Consideration of Intermediate Process Data

In formulating problem and approach, the research team originally intended to include a number of features representing intermediate process data – for example, characterizations of the variability of select process parameters over time, concentrations of metabolites at specific days in the process, intermediate versions of the final performance indicators, and outputs of earlier steps in the manufacturing process.

However, in the context of the desired learnings and final application of this model – that is, the ability to predict commercial-scale performance of a new program based off of small-scale data – these intermediate indicators would not be available in the test set. In this case, if that set of intermediate data was needed as an input to the algorithm, data from the select few clinical or toxicology runs would have to be utilized to impute the values, and thus assumed to be representative of all commercial-scale manufacturing moving forward (in spite of the fact that it may have been completed using a different process), likely confounding the predictions. Additionally, given the challenges of maintaining a reasonable ratio between observations and features and the impact these intermediate indicators would have on feature proliferation, the intermediate indicators are excluded in favor of focusing on only initial process parameters.

3.3.2 Selection of Engineering Features

A key area of investigation in this research is the impact of equipment engineering features on cross-scale predictive performance. Thus, it is necessary to specify exactly which engineering features are most important to include from a large set of possible features. These possible features include dimensions and specifications of the bioreactor, along with calculated values that characterize the fluid dynamics, forces, stresses, power dynamics, gas availability and transfer, and relevant ongoing biological processes.

A subset of 36 of these features is selected based on discussion with process development scientists, review of Amgen policies, procedures, and historical data, and literature review. The primary source of information on these mechanistic models and characterization of the bioreactor is Perry's Chemical Engineering Handbook, along with feature-specific literature review and research. However, this is supplemented significantly by conversations with process development scientists, process monitoring experts, facilities and engineering personnel, technology transfer specialists, and manufacturing personnel, along with review of internal models. The features are selected primarily based on perceived relative importance of the feature on the process performance and existence of the proper data and information to calculate it accurately at all scales of production under study, although interest and level of understanding are also considered.

Given the common factors and constants from which these features are calculated, collinear relationships may exist between the included engineering features. With the purpose of creating a comparison to the primary models which incorporated different explanatory variables via dimensionality reduction, partial least squares (PLS) regressions are added to the set of models to investigate comparative performance. Since this type of regression reduces dimensionality, it can be more difficult to interpret but is very adept at managing collinearity. It is seen in these analyses that these PLS regressions do not perform better than the primary models under investigation, and thus the researcher deems collinearity not to be a primary concern with the selected data and modeling methods. However, further exploration of model performance incorporating select subsets of engineering features to reduce potential collinear relationships could offer additional insight on this topic.

3.4 Predictive Hybrid Modeling: Architecture and High-Level Approach

In order to answer the stated research questions, a custom set of models is created with the purpose of investigating and drawing conclusions about the relationships between the selected process parameters and performance indicators of the production bioreactor across scales. This model also incorporates the inclusion of calculated equipment engineering features, serving here as mini-mechanistic models of the internal environment of the bioreactor, as an additional set of inputs for the purpose of determining their comparative importance and impact on predictive performance. A number of machine learning models are included, with each serving as its own competing model in any given analysis for the purpose of identifying the top-performing models for the problem at hand, along with discerning comparative importance of the features using cross-model feature selection analysis.

3.4.1 Architecture

The inputs to each model are the historical process data for the two processes and two scales being studied, along with calculated equipment engineering features for each scale. The outputs for each model are the selected performance indicators. For each analysis, out-of-sample performance and feature selection are evaluated for each model, and cross-model analysis of feature selection is performed. The models run for any given analysis include 10-16 different machine learning models and optimization methods. A visual depiction of this architecture is displayed below in Figure 4.

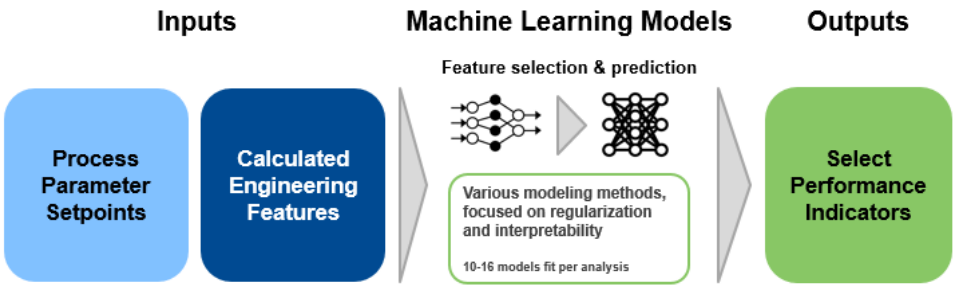

Figure 4. High-Level Machine Learning Model Architecture. Visual depiction of machine learning architecture utilized throughout all scenarios and analyses.

3.4.2 High-Level Approach

The project approach is focused on the creation of the aforementioned custom set of models, which are created to learn about process-specific and scale-specific relationships from historical process data, evaluate feature selection and predictive performance across several algorithmic techniques, and investigate the specific research questions posed in Chapter 1. This is accomplished by varying the subsets of data included in training and test sets, adding and subtracting equipment engineering features to investigate their impact on predictive performance, and analyzing the importance of each included feature. Four scenarios are constructed to test the stated hypotheses and ensure fair bases for comparison are created. These scenarios, and methods to accomplish each of these tasks, are explored in further detail in Chapter 4.

Chapter 4: Research Methodology

With the data, explanatory variables, response variables, and architecture defined, we now turn to the development of the predictive models. This chapter details the raw data collection, data pre-processing, and engineering feature creation and inclusion to provide a full picture of the data set used in each analysis. It then covers the machine learning techniques in use, discusses a novel cross-model approach to feature importance analysis, and concludes with a discussion of special considerations and a description of all scenarios we create for testing these models and the research questions posed.

4.1.1 Process Data Collection

Process data necessary for this analysis is housed in different places and formats depending on the scale, process, and a number of IT systems such as different electronic lab notebooks, especially given that the relevant lab data is collected over various historical periods. Thus, gathering the exact data necessary and associating it properly with other data sources to compile a single, final, uniform dataset on which to perform the analyses is a challenging, iterative, and time-consuming process.

4.1.1.1 Commercial-Scale Data

For commercial-scale data, data collection is fairly straightforward – with the help of several members of the Amgen process monitoring team, custom queries have been written to pull the relevant data and format it appropriately. The main challenges in this piece of the process centers on identifying the right data, ensuring consistency across processes and scales, and determining how to combine the much more robust commercial datasets with the small-scale datasets.

4.1.1.2 Small-Scale Data

Gathering the small-scale process data is difficult, mainly because the analysis requires going back to the initial lab notebook entries (as opposed to process characterization reports and other related documentation, which are complete and easy to access) and pulling together specific pieces of data from hundreds of individual Excel workbooks. In order to accomplish this, Excel macros and custom Python scripts have been written and leveraged to extract the relevant data in

a scalable way. Significant time was invested in ensuring that the data being utilized is correct and consistent across entries and processes (for example, as related to the specific sample value being considered the "final value"), pulling together associated analytical results from other sources, and double-checking that the data is consistent.

4.1.2 Data Pre-Processing

Process data imported into the model for analysis is examined for shape and outliers, and then pre-processed through several steps to ensure it is clean, complete, and ready for regression analyses. This involves eliminating incomplete values, encoding categorical variables, adding interaction terms, scaling the data, and splitting the data into training and test sets. Each step is covered in more detail in the following subsections.

4.1.2.1 Examining the Data

The full data set is first plotted using box-and-whisker plots to investigate the shape of each feature and response variable and identify outliers in order to inform later processing methods. This also serves as a data validation method, and helps the researcher identify terminated experiments and other anomalies early on in the project. The resulting plots are shown in Figure 5 and Figure 6 for the controllable process parameters for Process 1 and Process 2, respectively and Figure 7 for the response variables.

For the controllable process parameters, the box-and-whisker plot gives an indication of which parameters were tested most rigorously during the design of experiment (DOE), the relative ranges, and areas where certain process parameters had been tested outside their normal range (often in a challenge study, which is generally used to explore the possibility of expanding the range of a process parameter beyond what would normally be tested, as discussed in Chapter 2). A few key patterns arise. Certain process parameters were never varied – such as Process Parameter 2 – and a challenge study for Process Parameter 4 is clear in Process 2 only. All other parameters include statistical outliers, which in this case is a positive indication of the variability captured by the DOE process across these parameters. It is notable that this variation is on different scales; given these variables are not yet standardized, this is to be expected: the range of variation on a process parameter such as pH (where a variation of +/- 0.05 is significant) will be much

smaller than for other variables. It is also worth noting that Process Parameter 7 varies significantly in both sets of process data; this is as a result of scale differences in bioreactor configuration.

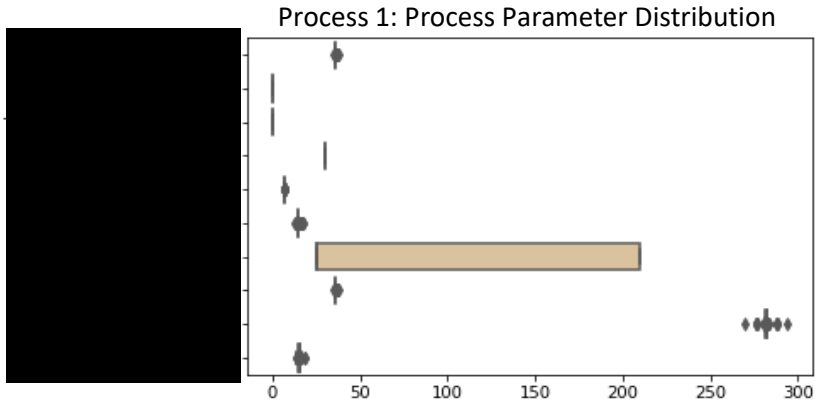

Figure 5. Box-and-Whisker Plot of Controllable Process Parameters for Process 1. Plot shows significant design variation in most selected process parameters, with the exception of Process Parameters 2, 3, and 4. Wide range of Process Parameter 7 attributed to scale-dependent bioreactor configuration.

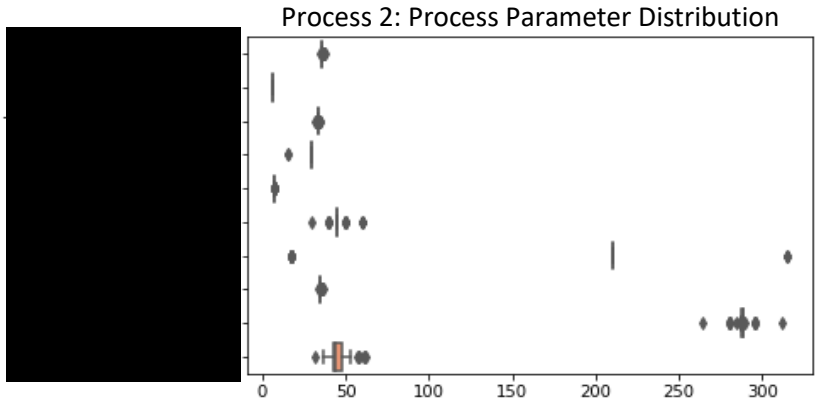

Figure 6. Box-and-Whisker Plot of Controllable Process Parameters for Process 2. Plot shows significant design variation in most selected process parameters, with the exception of Process Parameter 2; challenge studies are visible for Process Parameter 4, which is otherwise constant. Wide range of Process Parameter 7 attributed to scale-dependent bioreactor configuration.

For the response variables (combined for both processes given the same target outputs), box-and-whisker plots show that of the three selected performance indicators, only one has data points that would be considered statistical outliers. The interquartile ranges (IQRs) of the three performance indicators vary significantly: Performance Indicator 3's IQR is approximately 5x the width of Performance Indicator 1's IQR. The resulting plot is shown in Figure 7. The outliers noted for Performance Indicator 1 do not correlate directly to any of the challenge studies discussed

above, indicating that further study is necessary to understand the drivers of these irregular outputs. Additionally, it is clear that the selected performance indicators do show substantial variation, offering hope that models can be created to relate this variance to process parameters and equipment engineering features.

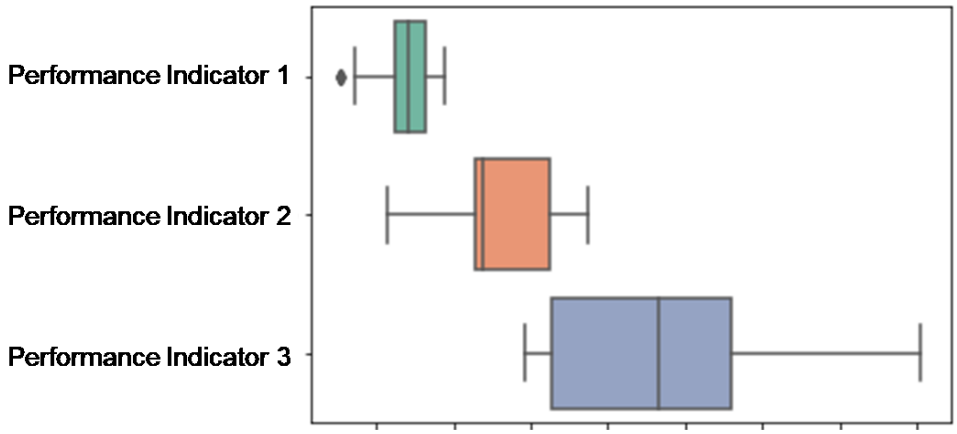

Figure 7. Box-and-Whisker Plot of Selected Response Variables. Only Performance Indicator 1 shows outliers, along with the smallest IQR. The IQR of Performance Indicator 3 is nearly 5x the width of that of Performance Indicator 1.

4.1.2.2 Treatment of Terminated Experiments and Null Values

The first step in data pre-processing is identifying and removing null values, since most of the algorithmic tools used in this project cannot take in data tables that contain these values. In this data set, null values arose from missing data (e.g., not every small-scale run has a full set of associated analytical results). In general, there are two strategies for managing null values – removing the related data (by row, or observation, or by column, or feature, based on a percentage of the total that is null), or imputing the data.

The first source of null values in this data set are from terminated runs, either in small-scale experiments or in commercial production, which do not have associated final performance indicator values. For these runs, it is clear that the proper approach is to remove these observations from the data.

The other source of null values in this data set is missing analytical results from a small number of small-scale experiments for which those product quality analyses were not done. Most often this type of data would be imputed (that is, estimated based on the data of other observations

most closely associated to it). However, in our case, imputing the data is not a sensible approach for two reasons: 1) the quantity of data is very small; and furthermore 2) each observation was designed to represent a specific set of conditions which vary in a designated way from the other observations. Thus, any attempt to impute the missing values would likely be both inaccurate and misleading for the machine learning models analyzing the data later. Given this context, any observations missing the response variable of interest are dropped from the dataset for the analysis of that specific response variable (but retained for analyses of the other response variables). This requires data pre-processing for each scenario and response variable, but ensures that all available data is used whenever possible.

For Performance Indicators 1 and 2, only three observations are excluded by this process, all of which represent terminated small-scale experiments for which there are incomplete data sets; this results in a full data set of 200 observations (121 from Process 1 and 79 from Process 2). For Performance Indicator 3, an additional 25 observations from Process 1 are excluded due to unavailable analytical data for that particular process and set of experiments, resulting in a full data set of 175 observations.

4.1.2.3 Treatment of Categorical Variables

The next step in data pre-processing is to convert categorical variables of value in the data set to a representative numerical form such that they can be included in regression analyses. In order to achieve this, one-hot encoding is used. The categorical variables of interest are encoded with "dummy variables" – that is, new variables are created for each of the possible values for the categorical variables, and observations are populated with 0s and 1s to indicate to which categorical value each corresponds.

The reason for selecting this particular type of encoding is that it is not prone to the same erroneous assumptions of numerical relationships between the categorical variables as sequential numerical-based encoding methods. Such methods might encode each possible categorical value as a number (for example, "apple" is 0, "orange" is 1, and "banana" is 2), which can indicate that there is an inherent sequential hierarchy or numerical relationship between them in a regression analysis. The drawback to our one-hot encoding approach is that it does add additional variables, which can be problematic when the matrix shape is suboptimal (that is, when there are too many

features for the number of observations). The proliferation of features is a concern in our case; however, since there are only three categorical variables included in the data set and only one of them is used in regression analyses (a process tag to differentiate between the two different processes included), one-hot encoding remains the optimal choice. The two additional categorical variables included in the data set are the batch number (the unique identifier of each observation) and the equipment tag (indicating scale, equipment, and process), both of which are used in data splitting, pre-processing, and adding in calculated engineering features before being dropped from the data set prior to regression analyses.

4.1.2.4 Interaction Terms

For all machine learning models except neural networks, polynomial interaction terms are included for each of the features representing process parameters. Second-degree polynomial interactions are selected by default for all analyses.

These polynomial interaction terms are not applied to engineering features when included in analyses, primarily for two reasons. First, the engineering features themselves are already intended to capture all the relevant interaction terms for the physical and chemical factors under consideration. Second, when engineering features are included in the creation of additional interaction terms, the shape of the data set becomes untenable for analysis given the limited number of observations available (meaning there are fewer than two observations per feature). We seek to maintain an approximately 4:1 or better ratio between the number of observations and number of features included in the data set for all models.

4.1.2.5 Scaling the Data

The next step in pre-processing is to scale the data. This is an important step given that the process parameters vary widely in scale (from small fractions to thousands). The scaling is necessary in order to draw conclusions about relative feature importance, control scale of coefficients (in particular for non-regularized regression, given the shape of the data set), and improve machine learning performance. This is considered best practice in most machine learning applications, since it is well-documented that many such algorithms (including linear regressions and neural networks) demonstrate better performance, optimization, and faster convergence when features are on similar scales and roughly normally distributed.[17]

Scaling is done using the Python scikit-learn StandardScaler package, which standardizes the data by subtracting the mean and then scaling to unit variance (dividing all values by the standard deviation), resulting in a distribution for each feature with a mean of 0 and a standard deviation and variance of 1. This is an ideal distribution for many machine learning models; however, it does have the drawback of distorting relative distances between features.

It is important to scale features for both the training and test set in a similar way; however, care must be taken to ensure that information from the test set does not leak into the training set during pre-processing and scaling of the data. This is accomplished by fitting the scaler to the training set only, and then applying the same scaler to both the training and test sets.

Finally, for optimal performance, response variables are scaled by simply dividing by factors of 10 until all are between 0 and 10; this preserves all of the information, relative variance, and relationships, but ensures that the features and response variablesaere on the same order of magnitude. This is because decreasing the spread of values limits error gradient values, lending more stability to the learning process.[18]

4.1.2.6 Training / Test Set Split

The data is then split into training and test sets prior to running any of the analyses. This is standard practice in order to ensure evaluation of model performance occurs on an out-of-sample test set in an attempt to avoid rewarding overfitting. Although it is common to split the data into three sets – training, validation, and test – when there are hyperparameter optimizations taking place, here we decide that the training/test set split is preferable in this scenario mainly because of the limited number of observations available (such that separating out an additional set of the data for validation is suboptimal). Given that most optimizations involve a limited number of hyperparameters, we believe that the training/test split is adequate in this setting, particularly in the context of a bias toward regularized regression methods with both L1 and L2 penalties that thus themselves penalize number of variables in such a way as to also combat excessive overfitting to the training set. A default proportion of 80% training data and 20% test data is utilized throughout the analyses.

4.1.2.6.1 Special Considerations for Training / Test Splits by Scenario

In certain scenarios, it is not possible to use a normal proportionality-based train/test split methodology. This is because the scenarios are designed with specific data subsets allocated to training or test sets. Due to the limited quantity of observations within each subset, in these scenarios it is necessary to allocate entire subsets of data manually to a training or test set. For example, when scenarios require all small-scale data and only a small sample of commercial-scale data to be in the training set, a custom function allocates that data appropriately into the training set, and then puts the remainder of the commercial-scale data into the test set.

In order to accomplish this type of prescriptive data allocation into training and test sets by scenario, a custom function is used to split the pre-processed data set into four subsets, each representing one unique process and scale combination (small-scale Process 1, commercial-scale Process 2, small-scale Process 1, and commercial-scale Process 2). A separate custom function then checks the pre-set allocation of data subsets associated with the scenario under analysis. For each subset, the function checks whether it is to be included in both the training and test sets; if so, that subset is split into training and test sets per the normal ratio and function (with the standard 80%/20% split) and added to the appropriate sets. For subsets allocated to only the training or test set, the subset is added in its entirety to the appropriate set. This enables maximal use of the data available, while testing a variety of scenarios and maintaining clear separation between training and test sets throughout the analyses.

4.1.2.7 Special Considerations for Neural Network Pre-Processing

Different pre-processing is used for neural network analyses, given that scaling the data and using interactions terms should not be necessary by nature of the technique itself, which takes on the responsibility of creating useful mathematical combinations of the provided inputs in the process of creating and tuning the network. Note, however, that both scaled and unscaled data sets are tested within the neural network function itself in order to test this assumption, and to ensure optimized performance within this context. Thus, to pre-process data for the neural networks, only the steps of dropping terminated runs and other missing values, encoding the process-identifying categorical variable, removal of specific features not desired in the analysis (such as the equipment

tag explicitly identifying the scale and any other user-specified features to be excluded), and inclusion of engineering features (where desired based on scenario) are performed.

Additionally, a specific train/test split is not used for the neural network analyses. Instead, K-fold cross-validation with a default of value of K = 10 (indicating 10 splits) is used to train, test, and evaluate the neural network models. This is because the size of the data set is particularly small for a neural network application, and thus K-fold validation enables maximal use of every observation while maintaining protection against rewarding excessive overfitting and evaluating the model on an out-of-sample set. Additionally, because of the limited dataset, lack of interpretability, and length of training time, the neural network was not one of the primary models analyzed or presented to Amgen stakeholders in results; it was included primarily for testing and to create model architecture that could be useful in the future with larger datasets (once available).

4.1.3 Addition of Engineering Features

Engineering features are calculated and included in the data set for certain analyses in order to evaluate whether their inclusion leads to improved predictive performance when algorithmic techniques are held constant.

4.1.3.1 Calculation of Engineering Features

Equipment engineering features selected for inclusion are calculated in a separate Excel model and worksheet for the purposes of accessibility, transferability, and use outside of this particular machine learning application. Calculation methods are based on proven equations for the physical, chemical, and biological factors of interest; additionally, the methodology was discussed and validated with experienced process development scientists from Amgen Inc.

The engineering features are calculated for each scale (including 2L, 3L, and 15,000L) and each process, since differences in the process parameters, shifts, and timing of the two processes vary and thus affect values for the calculated engineering features.

For values that varied over the time period of the production bioreactor unit operation (due to time-dependent changes in process parameter setpoints), weighted averages are calculated to arrive at a representative value for the entire operation for that process and scale. This is done to

enable comparability between processes and scales, without further proliferation of number of features included (to maintain the desired 4:1 or better observation to feature ratio discussed earlier).

4.1.3.2 Method of Inclusion in Data Set

As the engineering features are calculated based on process and equipment, unique equipment tags identifying both the scale and process are created and associated with each set of values. These same equipment tags are associated with each observation in the data set. When engineering features are desired in a given analysis, the full set of values are added to the data set based on these equipment tags using the Pandas 'merge' function during data pre-processing, and the full, augmented data set is used for all further analyses in that scenario.

4.1.4 Machine Learning Approach

This project takes the approach of using supervised machine learning models to discern relationships between the explanatory variables (process parameters, equipment and scale, and equipment engineering features) and selected performance indicators for the production bioreactor operation. Supervised machine learning is selected due to the quantity of data available, and then specific models are selected for inclusion, with a particular bias towards regularized linear regression models and ensemble of trees models. The primary reasons for these selections are the importance of feature selection, performance on a wide data set (with many features relative to the number of observations), and interpretability. Simplicity and known performance in regression analyses are also considered in selecting models for inclusion. Two neural network models are also included to test performance and evaluate applicability to the size, shape, and nature of this data set.

4.1.4.1 Review of Selected Algorithmic Techniques

For this application, the models to be included in the analysis are selected primarily for their ability to not only predict, but also to perform feature selection and deliver interpretable results. For this reason, linear regularized methods feature prominently in the analysis. As linear methods, these models represent the response variable as a linear combination of the explanatory variables. However, more specifically, these methods attempt to address the classical bias-variance tradeoff

by managing the complexity of the model via regularization so as to minimize both bias and variance and avoid overfitting to the training set. Specifically, these models penalize the complexity of the fitted model in different ways: 1) an L1 penalty term, which penalizes the number of variables, incentivizing sparse models with few variables and thus specializing in feature selection, and/or 2) an L2 penalty term, which penalizes the magnitude of the coefficients, thus incentivizing models with small coefficients (decreasing model complexity in this manner) but allowing the inclusion of several potentially collinear variables.[19] A key weakness of the L1 penalty method in a data set with multiple collinear variables (as is the case in this analysis) is that it will often select only one of a family of collinear variables (potentially arbitrarily and/or purely dependent on the hyperparameter value). The L2 penalty avoids this pitfall, but does not provide efficient feature selection, instead maintaining many variables with small coefficients. Thus, a method utilizing a convex combination of the L1 and L2 penalty terms is also included in an attempt to leverage the benefits of both methods.

The primary models deployed of this type are Lasso, which uses the L1 penalty; Ridge, which uses the L2 penalty; and Elastic Net, which provides the convex combination of both methods. All methods utilize Scikit-learn packages. Each of these methods requires one to two hyperparameter values; in order to obtain optimal performance, various iterations of each model are included and run for each scenario, with each version utilizing a different method to select the optimal hyperparameter. These include the base packages with default values, along with packages featuring cross-validation and Least Angle Regression (LARS) cross-validation, and Bayes-optimized hyperparameter selection using the Python Scikit-Optimize ("skopt") BayesSearchCV package.[20]

Based on the same selection criteria, an ensemble of trees algorithm is included in the analysis as well. The basic ensemble of trees method enables the use of decision trees which include different subsets of features and voting systems to help create the optimal model. These enable effective, interpretable analyses of what combinations of explanatory variables are necessary to characterize the response variable. Further method options can include the incorporation of dynamic feedback to improve efficiency and performance, "pruning" (cutting off branches of trees not providing sufficient value) for effective feature selection and efficiency, and

more. The model included here is known for exceptional performance with small-to-medium-size structured datasets as used in this analysis, and is called Extreme Gradient Boosting, commonly referred to as "XGBoost." This is an optimized gradient boosting algorithm that leverages parallel processing, tree-pruning, and regularization to deliver efficient predictions and feature selection with highly-interpretable results (presented via a feature importance ranking).[21] Thus, this is a valuable complement to the selected linear regularized methods, and is implemented using the Python package xgboost.XGBRegressor.[22]

Finally, a three-layer neural network is constructed (one input layer, one hidden layer, and one output layer) using the Keras Deep Learning Library in Python.[23] Although the size of data in terms of number of observations included here is generally considered to be too small for effective neural network performance, and neural networks are not the primary method of interest for reasons of feature selection and interpretability, it is included as an interesting comparison point in order to investigate neural network performance with the current data set, and to create the infrastructure for further study once the data set is expanded. Interaction terms are not included in the data set for the neural network given that the network should be capable of combining the initial input terms appropriately on its own, but it is tested both with and without engineering features.

In addition, partial least squares regression (PLS) models are run for each scenario as well for use primarily as a comparison point, and to provide a potential warning sign that a sub-optimal group of explanatory variables is being included (particularly with regard to engineering features). This is because PLS models aim to reduce dimensionality of the problem by creating new combinations of the explanatory variables, often succeeding in making predictions with fewer variables, but suffering from a lack of easy interpretability since the initial explanatory variables do not match up directly with those use in the fitted model. Thus, while PLS is not an optimal choice for the primary analysis given the importance of feature selection and interpretability in our case, it provides a useful comparison in terms of performance. If PLS is regularly out-performing the other models it could indicate that the most helpful explanatory variables (including linear combinations and interaction terms) are not being included in the model.

4.1.4.2 Performance Assessment and Metrics

Models are compared and evaluated on the basis of out-of-sample R^2 (as evaluated on the test set) and root mean squared error (RMSE) values on the prediction set. The primary metric for performance improvement is an increase in the maximum R^2 value from the top-performing models included in the scenario. Secondary metrics are decreases in RMSE and/or decreases in the standard deviation of the maximum R^2 values as determined by K-fold cross-validation.

4.1.5 Feature Importance Analysis

In this effort, an evaluation of the most important features for predicting each of the selected performance indicators (the response variables) is a key outcome of the project. Given this goal, it is important to consider and develop a robust approach to feature selection and analysis.

4.1.5.1 Cross-Model Importance Analysis

For this feature importance analysis, we consider a more robust approach to feature selection than examining a single top-performing model. Instead, a cross-model ensemble view of feature selection is created, by comparing frequency of each feature being selected by multiple models. Based on user inputs, this analysis can consider all models or only a top-performing subset of models as evaluated by R^2, and the number of models can be specified by the user (with a default of three). In order to complete this analysis across different types of models, it is necessary to create a way of extracting equivalent feature selection information across all methods. In particular, connecting feature selection outputs across regularized regression and ensemble of trees techniques is of interest. The feature selection properties of these two categories are discussed in the following two subsections, followed by a description of the cross-model technique we create to connect them.

4.1.5.2 Regularized Regression Models

For regularized regression models (such as Lasso), there is no built-in functionality to create a feature importance plot or determine the percentage of variability explained by a given feature's inclusion. However, in this case, since the data has been scaled and the regularization methods penalize complexity, thus forcing feature selection via non-zero coefficients, there is an opportunity to consider these non-zero coefficients as indicative of feature importance, particularly

when considered across top-performing models. In order to decrease the risk of a model "selecting" nearly all features with coefficients very near to 0 (particularly in the case of Ridge analyses), a threshold with a default value of 0.05 is established, and features are required to have a coefficient above this threshold in order to be considered "selected" by this feature importance analysis.

4.1.5.3 Ensemble of Trees Models

For ensemble of trees models (such as Extreme Gradient Boosting, or XGBoost), the concept of feature importance is central to the analysis itself, and thus extracting the information is much simpler. The feature importances in this case are given as a number between 0 and 1, representing the relative importance of each feature in constructing the final boosted decision tree; the more often the given feature is used to make key decisions in each decision tree, the higher its relative importance. Since these feature importances are relative to each other, it enables a natural comparison and ranking of the most important features. For these models, feature importances are extracted using the built-in features of the relevant Python packages. In a similar fashion to the treatment of regularized regression coefficient-based importances, the threshold with a default value of 0.05 is applied to the importances, and only features with an importance level above the threshold are considered "selected".

4.1.5.4 Combined View

After completing the feature selection analyses of both the regularized regression and boosted trees models, lists of selected features for each model were compiled. Based on model R^2 values and the user-selected number of models to be included (anywhere from only the top-performing model to all models), the selected feature lists of only the selected models were compiled and used to create a frequency analysis of cross-model feature selection. For example, with the default value of selecting the top 3 models by performance (as evaluated by R2), lists of selected features for only the top 3 models would be extracted, and then each feature would be ranked on a frequency plot based on its selection in 0, 1, 2, or 3 of the top-performing models. A similar analysis is done by default for all models, since the differences can be enlightening – for example, when only 1/3rd of the models selected a single feature, but all of those were in the top-performing subset.

Thus, a novel method of cross-model feature selection analysis was used to draw additional conclusions about the importance of various features in scale-up analyses.

4.1.6 Challenges and Special Considerations in Pre-Processing

The multi-scenario analysis and complex data pre-processing and allocation introduce a few additional considerations worth discussing here. In particular, this section will discuss constraints on the inclusion of multi-scale data in all analysis focused on engineering feature impact, the rationale and methodology for commercial sampling in select scenarios, and the limitations on the use of K-fold cross-validation to determine robust R^2 values with standard deviations.

4.1.6.1 Scale Variability Required for Engineering Feature Analysis

In the creation of scenarios for testing, each of which includes different subsets of available data at each scale, engineering features, and algorithmic techniques, it is important to note that it is not possible to test the impact of engineering features on predictive performance without including a balanced subset of process data across both scales in the training set. The reason for this is that the engineering features themselves are constant for a given process and scale, and nearly constant across all data at each scale – for example, the size, shape, and set-up of the bioreactor (and thus resulting geometric ratios and power dynamics, before considering agitation speed and similar process parameters) will be constant across every process and run; the bioreactor itself is not changing, only the settings. Thus, if one were to test the impact of engineering features on the performance of a model with only small-scale data in the training set (or even a training set including a commercial sample, as will be discussed in the next section), it is nearly certain that none of the features would be assigned any importance by the model – given their constancy, they would in no way contribute to explaining the variability of the data in the training set.

As a result, all analyses specifically intended to illuminate the impact of engineering features split the full set of combined small-scale data and commercial-scale data into training and prediction sets according to the specified split discussed earlier (with a default split of 80% / 20%).

4.1.6.2 Commercial Sampling

In order to create fair comparisons between current-state linear regression analyses from process characterization and the new models, it is important to consider the quantity of

commercial-scale data available at the time of process characterization (for both past comparison and future application of this model to new programs). In general, before full scale-up, process data is available from initial commercial runs (used to create material for clinical and toxicological studies) and occasionally not-for-human-use, or "engineering" runs. These are used in the creation of "with scale" regression equations, which are used for the current-state baseline in this analysis.

In both current-state and comparative analyses, although the exact initial commercial data from process characterization is not always available, the effect is replicated by creating a "commercial sample" in a quantity proportionally representative of the commercial-scale data available at the time of process characterization. In order to do this, a randomly-selected representative quantity of commercial-scale data is removed from the prediction set and included in the training set.

4.1.6.3 Limits to Application of K-Fold Cross-Validation

It is worth noting that the cross-validated out-of-sample R^2 value and standard deviation of that value as determined by K-fold cross-validation is only applicable to scenarios in which a standard train/test split can be applied to the data. That is, scenarios in which data subsets have to be manually allocated to the training or test set (as discussed in Section 4.1.2.6.1) do not lend themselves to cross-validated out-of-sample assessment by this method. As such, for those scenarios, all R^2 values are taken as point values from performance on the test set.

4.1.7 Scenario Analysis

Four scenarios are constructed in order to test the stated hypotheses, ensure fair bases for comparison, and explore the future applicability of the current model to new processes.

Scenario #1 applies linear regression equations from process characterization (which includes both small-scale experiment data and limited at-scale data from clinical runs) directly to commercial-scale parameters to establish a current-state baseline performance. Scenario #2 uses the full subset of small-scale data and a similarly-sized sampling of at-scale data with more advanced models to determine if algorithmic techniques could improve predictive performance without the inclusion of any additional data. Scenario #3 examines the impact of hybrid modeling by incorporating calculated engineering features and the full dataset to evaluate the impact of

engineering features on predictive performance. Finally, Scenario #4 serves as an initial exploration into the application of transfer learning in this context (specifically, process-to-process transfer learning): the model is trained on all available data, including engineering features, for Process 1, along with the small-scale data and a commercial sample for Process 2, and then evaluated on predictions of commercial-scale performance for Process 2.

This final scenario acts as a preliminary exploration of whether this data could be used in a transfer learning context. Transfer learning is defined as "the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned." The idea is to "pre-train" a model on a base dataset and task, and then to repurpose or "transfer" these learned features to a second target network, dataset, and task. It is frequently used in deep learning tasks such as image recognition – where one can download a model pre-trained on the ImageNet database (used as the base dataset with the task of image classification) and re-purpose it; given the resource intensity required to train a model on an enormous image database, this can save days or weeks of training versus starting from scratch. However, another application of this technique is less focused on reducing training time and more focused on enabling the creation of advanced models for a task where limited data exists.[24] That is the scenario envisioned here – if an Amgen data scientist wants to model a brand new process in the scale-up process, they will have very limited data with which to work, and far too little to develop a sophisticated model. However, with transfer learning techniques a sophisticated model could be created based on a complete dataset of all historical process data (across all scales and products), and the learned features could be "transferred" to a new model re-purposed to predict performance of the new process at commercial scale. It is our belief that this would be the most powerful application of this technique in the biomanufacturing and scale-up process. However, it is worth noting that the proof-of-concept test in Scenario #4 is quite preliminary by nature of the small quantity of data and variability available for extrapolation across products and scales.

These scenarios are laid out in further detail in Table 1.

Table 1. Modeling scenarios analyzed in the evaluation of hypotheses and exploration of future applicability.

Scenario	Training Set	Prediction Set	Methods	Objective
1: Current-State Baseline	• Small-scale DOE data (~30 experiments / process) • Commercial sample	• Commercial-scale data	• Linear regression equation from process characterization	Assess performance of previously-determined regression equation on commercial-scale data
2: New Baseline – Advanced Algorithmic Model	• Small-scale data • Commercial sample	• Commercial-scale data	• Linear regression • Regularized linear regression • Ensemble of trees • Neural network	Determine if different algorithmic techniques improve model performance; establish new baseline
3: Hybrid Model	• Small-scale data • Commercial-scale data • Engineering features	• Small-scale data • Commercial-scale data • Engineering features	• Linear regression • Regularized linear regression • Ensemble of trees • Neural network	Determine if inclusion of engineering features improves model performance
4: Process-to-Process Transferability Testing	• Process 1 small + commercial-scale data • Process 2 small-scale data • Engineering features	• Process 2 commercial-scale data • Engineering features	• Linear regression • Regularized linear regression • Ensemble of trees • Neural network	Explore transferability of learned scaling relationships between similar processes (initial proof-of-concept only)

Chapter 5: Results and Discussion

This chapter presents the findings of each analysis, discussing the predictive performance we achieve in each scenario and implications for the related hypotheses. The chapter begins with a review of the model performance for the first two scenarios (the current-state and advanced algorithm baselines), and then turns to a discussion of scenario #3, in which we investigate the impact of the engineering features, review the results of the feature selection analysis, and consider the resulting engineering feature paradox. The chapter then reviews scenario #4 and the concept of process transferability before concluding with a discussion of the challenges and limitations of this study.

5.1 Model Performance

Models are evaluated individually on the basis of R^2 and RMSE, as discussed in Section 4.1.4.2. Model performance is found to vary significantly by response variable and data included (as determined by the scenario and inclusion or exclusion of the calculated engineering features). This provides an opportunity to not only discuss the results for each response variable, but also to examine the areas where the results align and diverge, to gain additional understanding of the factors that govern bioreactor performance.

5.1.1 Baseline Assessment: Current-State and Advanced Algorithmic Baselines

The first part of this analysis is intended to analyze the current-state baseline (Scenario #1 in Table 1) by determining whether the regression equations calculated during process characterization can be applied directly to predict commercial-scale performance. Note again that these regression equations are not currently created for this purpose; however, it is an important test to understand if such an application is feasible. We then compare to a new baseline (Scenario #2 in Table 1) to understand the performance of a model using the same data set, but more advanced algorithmic techniques for the analysis. We continue to compare performance of these techniques with a multi-process data set, and finally with engineering features (Scenario #3 in Table 1).

Thus, five sub-scenarios are created and tested: 1) predictions using directly-applied regression equations based only on DOE data; 2) predictions using directly-applied regression equations "with scale," which are created in process characterization to test scale effects by including a subset of commercial-scale runs from clinical and toxicological studies; 3) predictions from the advanced algorithmic model trained on small-scale data with a commercial sample representative of that used in creating the scale adjustment in sub-scenario 2; 4) predictions from the same advanced algorithmic model, but this time with data combined from both processes; and 5) predictions from the same multi-process advanced algorithmic model, but this time with engineering features.

Coefficients for the first two scenarios – representing process characterization regression models – are hard-coded directly from JMP regression models and process characterization reports. To ensure a fair comparison with these process characterization regression models, which consider only one process, the first three sub-scenarios consider data from only one process; the final two include combined process data, which is necessary to evaluate engineering features. Our expectation is that each progression will result in better predictive performance, with the caveat that introduction of data from both processes could have a negative impact due to increased variation.

The results of each of these sub-scenarios across both processes for Performance Indicator 2 is shown below in Figure 8, which plots the best R^2 value of each sub-scenario to display relative performance. As expected, performance generally improves with every advancement; the one exception is in the scale adjustment (indicating likely non-representative over-fitting in the scale adjustment during process characterization analysis). The impact of engineering features is marginal and quickly disappears with rounding. The other performance indicators display very similar results, although Performance Indicator 3 does show a slight dip in performance with the change to multi-process data in sub-scenario #4 due to the additional variation introduced and the fewer included observations due to missing data.

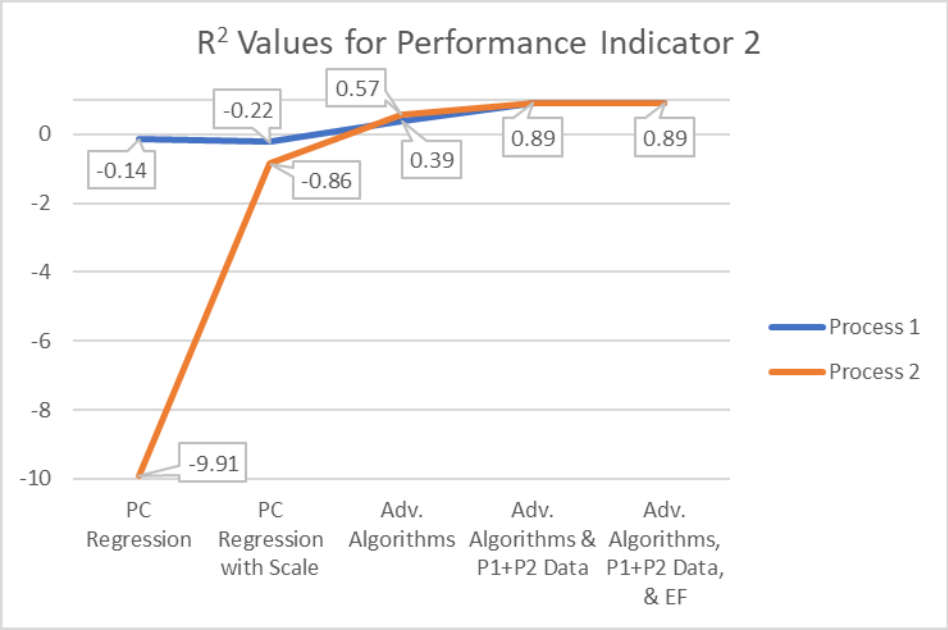

Figure 8. Predictive Performance of Process Characterization Regressions versus Advanced Models for Performance Indicator 2. Values shown are R^2 values for the top-performing model where multiple exist (in the advanced algorithm models). Trends representative of all processes and performance indicators. Legend: "PC Regression" represents process characterization regression equations applied directly to commercial setpoints; "PC Regression with Scale" indicates that the equations used are those with scale adjustment from the process characterization report; "Adv. Algorithms" represents the advanced algorithmic models with single-process data; "Adv.Algorithms & P1+P2 Data" represents the same model with combined process data; and "Adv.Algorithms, P1+P2 Data, & EF" represents the hybrid model in which calculated engineering features are included with the combined process data set.

There are several takeaways from these analyses. First, process characterization regression models cannot be applied directly to the prediction of commercial-scale performance. Even models with in-sample performance having R^2 of 0.91 (for Process 1, Performance Indicator 2) fall short when asked to predict at a different scale, dropping to an R^2 of -0.14. Scale-adjusted process characterization models do significantly improve performance on a commercial-scale data set, as would be expected; however, this improvement is insufficient to result in a positive R^2 value. In fact, all R^2 values from process characterization regression models (including those with scale adjustments) are negative across response variables and processes. This indicates that predicting the mean value across all observations would result in better performance (this is represented by an R^2 value of 0). It is notable that the mean is a difficult standard to beat in this case, since biomanufacturing processes generally operate within such tight controls and have minimum

variability, particularly at commercial-scale. However, a predictive model will provide value only if its predictions are able to beat this standard.

Second, more advanced algorithmic techniques can offer a significant improvement in performance when provided the same initial training set and test set. This is shown by the jump from negative to positive R^2 values for both processes in Figure 8, with improvements of +0.79 and +1.25 (due to the sign change) for Process 1 and Process 2, respectively. The R^2 values are not particularly impressive at less than 0.6; however, it is clear that the techniques themselves provide additional value.

Third, performance improves across performance indicators when the data sets for both processes are combined at each scale and the same analysis is performed. This finding emphasizes the importance of the quantity of data, but also lends credence to the idea that these kinds of models can learn from data coming from similar processes, without them needing to be identical. This is important for future development of the algorithm.

An additional takeaway arises from the impressive performance of the models for Performance Indicator 2. With advanced algorithmic approaches and a combined data set, the XGBoost model achieves an R^2 of 0.89, and even matches in-sample performance with an R^2 of 0.91 with the addition of engineering features. In-sample performance in this case is represented by performance of the process characterization regression model with scale on only the single-process small-scale data set on which it was trained. This indicates that fairly limited data (in terms of richness and quantity) can enable high-fidelity predictions of Performance Indicator 2, and suggests that the small-scale process is representative of the commercial-scale process in terms of the relationship of controlled parameters to output for this performance indicator.

Finally, this analysis validates the need for a better predictive model that can translate, predict, and analyze results across scales, as this project intends to create.

5.1.2 Best-Performing Models with Full Dataset Before Engineering Feature Addition

For this analysis, the full set of small-scale and commercial-scale data is included in order to enable a fair comparison of pre- and post-engineering feature inclusion (which requires a mixed training set of both scales, as discussed in Section 4.1.6.1).

Overall performance demonstrates a clear and significant improvement over both baselines, with top R^2 values for Performance Indicators 1, 2, and 3 of 0.790, 0.902, and 0.888, respectively, among the various model performances for each response variable. It is worth noting that for Performance Indicator 2 and 3, these top-performing models match and out-perform the current-state baseline, respectively. This baseline is the regression focused entirely on small-scale data from the relevant process characterization report, without the added complication of multi-scale data sets and prediction tasks.

The best-performing model in the dataset without engineering features for all three response variables is Bayes-Optimized Ridge Regression, which out-performs the second-best model by R^2 values of 0.02 to 0.05 and others by much more for Performance Indicators 1 and 2.

This indicates that the L2 penalty method, in which the penalty term of the loss function is the squared magnitude of the coefficient, proves particularly useful in the analysis with fewer features (in which variable reduction was less critical). This changes in the dataset augmented by engineering features, where the feature selection is much more critical, as will be discussed in the next section.

An additional takeaway is that Bayes optimization outperforms all other hyperparameter selection methods studied here, including default values, cross-validation, least-angle regression of squares (LARS), and the combination of these methods. This optimization method appears to be particularly powerful for Elastic Net, bringing improvements across all response variables, both with and without engineering features, of 0.11 to 0.77 in R^2 performance as compared to the Elastic Net baseline algorithmic package.

5.1.3 Impact of Engineering Features

The addition of engineering features does not display the universally positive impact that was hypothesized. Although inclusion of these features does tighten the distribution of R^2 values, the impact is variable in magnitude and sign for each response variable and algorithm. A selection of R^2 values by model before and after engineering feature inclusion is shown below in Table 2. Each response variable is represented by its abbreviated name – for example, Performance Indicator 1 is represented as PI1. This analysis reveals that some algorithms benefit more than others from the engineering features, that Performance Indicator 1 appears harder to predict than the other response variables, and that only Performance Indicator 3 benefits across all models from adding engineering features.

Table 2. R^2 Values and Impact of Engineering Features by Algorithm and Performance Indicator.

	Without Eng. Features			With Eng. Features			Eng. Feature Impact		
	PI 1	**PI 2**	**PI 3**	**PI 1**	**PI 2**	**PI 3**	**PI 1**	**PI 2**	**PI 3**
Lasso Bayes-Optimized	0.67	0.85	0.87	0.73	0.86	0.89	+0.06	+0.02	+0.02
Elastic Net Bayes-Optimized	0.73	0.85	0.87	0.72	0.83	0.89	-0.01	-0.02	+0.01
Ridge Bayes-Optimized	0.79	0.90	0.88	0.72	0.85	0.89	-0.07	-0.06	+0.01
Lasso	0.73	0.87	0.87	0.75	0.85	0.89	+0.02	-0.02	+0.01
LassoCV	0.73	0.86	0.87	0.71	0.83	0.89	-0.02	-0.03	+0.01
XGB	0.73	0.88	0.82	0.76	0.89	0.84	+0.03	+0.01	+0.01
Elastic Net	-0.07	0.57	0.62	-0.07	0.57	0.78	+0.00	+0.00	+0.16
Best Model Performance	**0.79**	**0.90**	**0.88**	**0.76**	**0.89**	**0.89**	**-0.03**	**-0.01**	**+0.01**

This analysis reveals several new pieces of information. First, Performance Indicator 1 predictions perform the worst of the response variables across nearly all models, indicating that other factors not included in this analysis may be necessary for high-accuracy predictions.

In terms of performance, it is interesting to note that the top-performing algorithms often changes between the pre- and post-engineering feature addition. Although it is dominant in the data set without engineering features, Bayes-Optimized Ridge Regression is replaced by Xtreme Gradient Boosting (XGBoost) in top-performance for Performance Indicators 1 and 2 once engineering features are included. Bayes-Optimized Ridge Regression retains the best performance for Performance Indicator 3, but essentially ties with several other models (edging out Bayes-Optimized Elastic Net by only $6.5*10^{-6}$).

XGBoost is the most improved on a per-algorithm basis with the addition of engineering features (providing a larger data and feature set). Only Bayes-Optimized Lasso shows a similar consistency in performance improvement across response variables. Thus, the relatively more aggressive feature selection algorithms, including the L1 penalty used by Lasso that forces coefficients to 0 to avoid a penalty for their inclusion, and the ensemble-of-trees-based methods that prune branches, perform better with a richer set of features from which to select. This appears to indicate the presence of valuable information in the engineering feature dataset not previously existent in the setpoint data.

It is also worth discussing the differential impact of the engineering feature inclusion on each response variable. Performance Indicator 2 is the least impacted. In fact, most algorithms perform worse when predicting this variable with the additional data. We see relatively mixed results for Performance Indicator 1, which also appears the most difficult to predict – predictions for this response variable perform the worst across nearly all models in both cases, indicating that additional data not included in this analysis may be necessary for improved performance. Performance Indicator 3 is the only response variable for which every method improves with the additional features, with an average impact of +0.036 and an increase in top performance of +0.013. This impact on the top R^2 value for each performance indicator can be seen visually below in Figure 9.

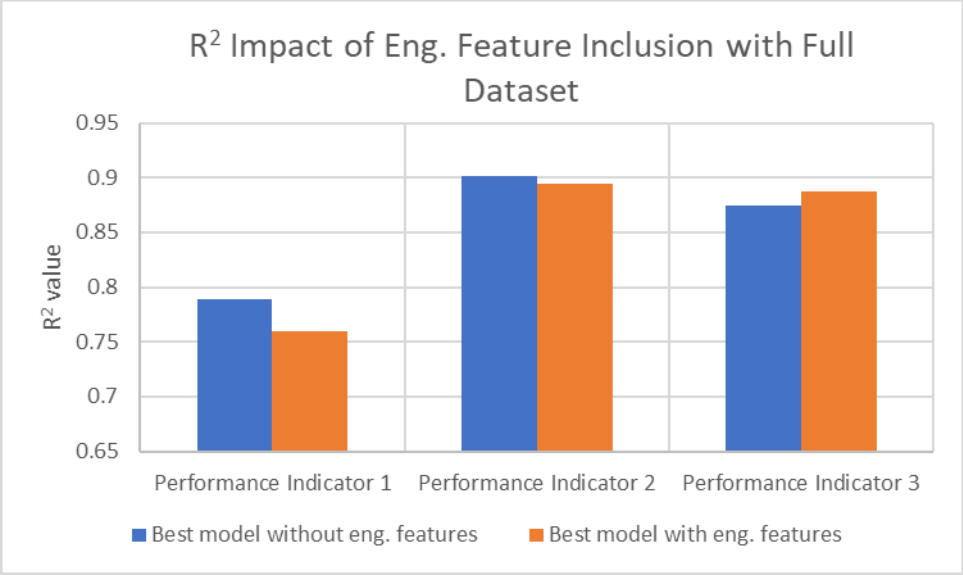

Figure 9. R^2 Impact of Engineering Features by Response Variable. Comparison of R^2 value of top-performing model for each Performance Indicator with and without engineering features included in the dataset. Performance Indicator 3 is the only response variable to show improvement in the best-performing model from the additional information.

This lends credence to the hypothesis that a portion of the improvement in performance for Performance Indicator 3 could be due to the increase in the richness of a dataset more limited in observations and variability from small-scale experiments. It is noteworthy that engineering features could augment the data set in such a way without any direct characterization of the process parameter variability, and indicates that there may be hope for an expanded version of this approach to enable Amgen to do more limited small-scale challenge-studies and experiments in the future for new processes, significantly decreasing the time and resources required for early process development and characterization.

It is also informative to represent the impact of engineering features visually. The pair of plots in Figure 10 represent the predicted versus observed values for an XGBoost algorithm before and after the inclusion of engineering features. These plots represent an improvement of +0.03 in R^2 performance after the addition of engineering features. It is notable that while overall performance improves, it appears that the impact differs between the two scales – predictions for commercial-scale performance, as represented by the tighter cluster in the lower left corner, worsen slightly while predictions for small-scale performance in the top right portion of the plot improve. This is a demonstration of the XGBoost model using engineering feature information to

enable better cross-scale prediction in select scenarios (these visual plots show minimal changes in scenarios where the R^2 performance remains approximately the same).

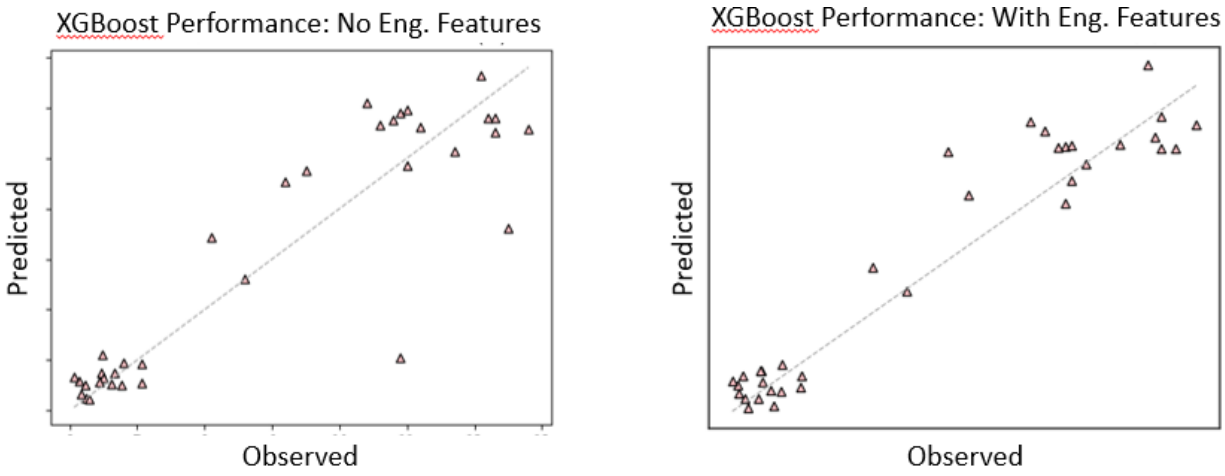

Figure 10. Visual Impact of Engineering Features: Predicted vs Observed. These plots represent the predicted versus observed values for an XGBoost model; the plot on the left represents a model using a data set with process parameters only, whereas the plot on the right represents a model using a data set with engineering features added. The plot on the right represents an improvement in overall R^2 performance of +0.03. Differential impacts on the two clusters appear, showing slightly worse performance in the bottom left and improved performance in the top right.

Thus, it appears that the impact of engineering features on model performance is quite variable but relatively insignificant with the current size and quantity of data. However, it is also clear that these features can have a positive impact – particularly for more restricted data sets, and for more aggressive feature-selection algorithms – and that this improvement likely indicates the existence of valuable information not present in the setpoints. This is supported by the findings of the feature selection analysis, which will be discussed in the next section.

5.1.4 Feature Selection Analysis

Overall, the selected regularized regression techniques do an impressive job in performing feature selection on a dataset with more than 50 features. The minimum number of features selected is five, with most models using between five and ten features. The maximum number of features selected is 47, and it is notable that top performer Bayes-Optimized Ridge Regression is, unsurprisingly, often at the top of the list in terms of a larger number of features selected (as the penalty term encourages small-magnitude coefficients but not reduction of features). However,

XGBoost, the top-performer in the engineering feature-included datasets, often required five or fewer variables to account for 95%+ of its prediction.

It is also notable that Partial Least Squares regression (PLS) is never the top-performer in these studies. This is considered a positive indication, in that it suggests that the features included in the analysis are the right ones given the same basis; that is, there is no other combination of the provided information – two or more features combined via mathematical operation – that would enable better predictive performance. Since the engineering features are calculated values acting as unique combinations of the same basic information about the process and the physical bioreactor, this is an important comparison point. Had this not been the case, it would have been important to investigate the vectors created by PLS and seek to replicate them in order to create the feature set that would enable the best predictive performance by the other algorithms, and through down-selection of those features, allow the most learning about what is critical in the process.

5.1.4.1 Analysis of Commonly-Selected Features Across Regularized Methods

It is notable that the features selected vary significantly by performance indicator, across both process setpoints and engineering features. The commonly-selected process setpoints are unsurprising for any biomanufacturing professional: setpoints for pH, temperature, agitation, and seed density are selected most frequently, with squared values of pH and temperature setpoints appearing important for Performance Indicator 1, and pH and seed density being most important for Performance Indicator 2. In terms of calculated engineering features, the only engineering feature that is selected by at least one model across all three performance indicators is the D_i/D_v ratio for impeller 1, indicating that this dimensional ratio may be a good indicator of the size and environment of a bioreactor. Perhaps more interesting is that for the performance indicator most assisted by the inclusion of engineering features, Performance Indicator 3, this ratio is selected in concert with other geometric ratios of the vessel – specifically, the aspect ratio at test volume and raw vessel diameter – along with measures of shear stress and power. This indicates that geometrically-based calculated features may work well in predicting Performance Indicator 3, particularly in concert with measures of shear stress.

5.1.4.2 Selected Engineering Features by Response Variable

For Performance Indicator 1, measures of power (specifically Total Power and Power, Impeller 1) and D_i/D_v ratio for impeller 1 are selected by a minority of models. It is notable that the model that selected for all three is one of the top three models by performance, indicating that there is information contained that helped at least one model outperform many others. However, the best-performing and most-selective models focus only on pH, temperature, seed density, and agitation setpoints. In fact, the best-performing model is able to predict with an R^2 value of 0.708 using only the pH setpoint, pH^2, and $Temp^2$. This is also the response variable that appears to be the most difficult to predict overall with the information provided; it is possible that there are variables not included in this analysis that could better inform and improve performance, and in future work it is likely worth discussing with process development scientists what that may be.

Results for Performance Indicator 2 display similar patterns. It is notable that Performance Indicator 2 shows the least improvement on a by-algorithm basis from engineering features; when engineering features are added, all algorithms worsen in performance with the exception of XGBoost and Bayes-Optimized Lasso. The best-performing and most-selective models focus on the setpoints for pH and seed density for the vast majority of the necessary information. A representative feature importance plot from an XGBoost model with an R^2 value of 0.894 is shown in Figure 11; in this case, >90% of the importance is attributed to the initial viable cell density (VCD) setpoint and pH setpoint. Several models also select agitation and temperature setpoints and one engineering feature – the vessel diameter, a clear indicator of scale – as important. Two models also select the D_i/D_v ratio, but it does not appear to improve performance.

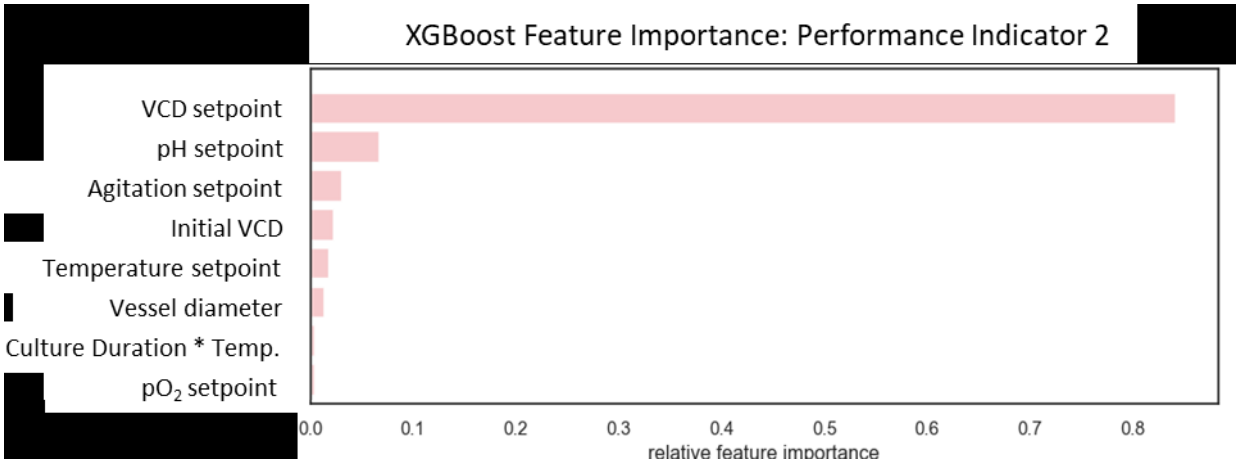

Figure 11. XGBoost Feature Importance for Performance Indicator 2. This feature importance plot shows the relative importance of each selected variable in predicting Performance Indicator 2, with engineering features included in the training and test sets. 90.8% of the relative importance is attributed to Process Parameters 6 and 5 (at 0.841 and 0.067 feature importance values).

This information, in concert with the slight decrease in the best R^2 value after adding engineering features, indicates that Performance Indicators 1 and 2 are better predicted by their setpoints across scales than with engineering characterizations of the environment, and points to the key sources of valuable information for each among their setpoints.

However, in the case of Performance Indicator 3, the results are quite different. Performance for every single algorithm improves with the inclusion of engineering features, and engineering features make up a significant proportion of the selected features in top-performing models. Furthermore, there is a clear correlation between the models which select more engineering features and improved performance.

Of the three top-performing models, all select the D_i/D_v ratio for impeller 1, max shear stress, and max shear rate. In the broader set of nine models fit for Performance Indicator 3, it is worth noting that seven models selected the D_i/D_v ratio, but only four or fewer models selected additional measures of shear. Among these, the top-performing models represent three out of four models that select these additional features – maximum shear stress, maximum shear rate, and an alternative maximum shear rate that considers the blade diameter of impeller 1. This appears to indicate a correlation between these engineering feature selections and improved performance. A similar pattern is seen for the power-focused engineering features. Only two of the nine models

select total power and power imparted by impeller 1; both are in the top-performing models, revealing another correlation between their selection and predictive performance.

These results indicate that the impeller geometry (in relation to vessel diameter), along with its imparted power and resulting shear, are the most important equipment engineering features for accurate predictions of Performance Indicator 3. Figure 12 shows the set of selected process parameters and engineering features by frequency of selection in the top three-performing models for Performance Indicator 3.

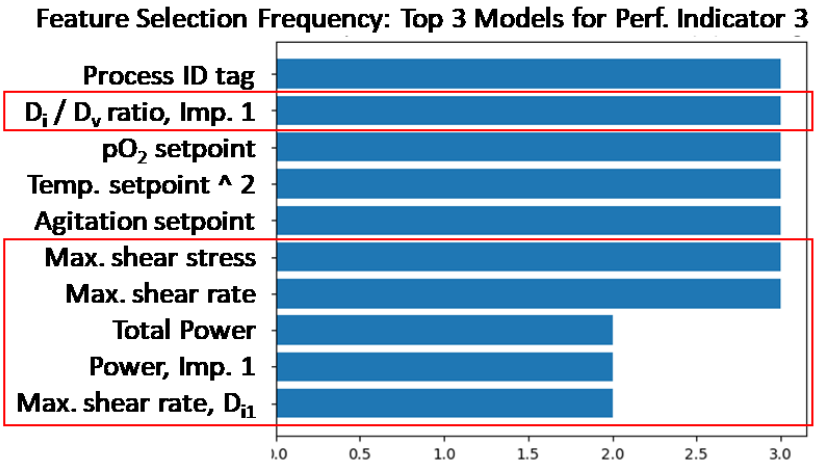

Figure 12. Cross-Model Feature Selection Analysis for Performance Indicator 3. The features outlined in red are the engineering features selected by the model. Maximum shear stress and shear rate are selected by only four models of the total set, of which three are top-performing, and power features are selected by only two models, both of which are top-performing as determined by R^2 performance. Thus, shear and power appear to be key contributors towards accurate commercial-scale predictions for Performance Indicator 3. Max. shear rate, D_{i1} represents the maximum shear rate considering the diameter of the blade on impeller 1.

One additional note on these cross-model feature selections is that it is quite interesting that the pO_2 setpoint is so commonly selected, considering that it has a constant value for all observations except for two small-scale challenge studies in one process. From this perspective, its inclusion is a powerful indicator of the importance of dissolved oxygen in determining the output of Performance Indicator 3.

It is also worth considering the relative importance placed on each of these selected features, which in this case is assumed to be represented by the magnitude of the coefficient in a fully normalized dataset. For this, discussion will focus on three models and their relative weights placed on selected features: the best performing and the most selective (among all and among the

top three models by performance), all of which have R^2 values of ≥0.84 for this performance indicator.

The most selective model out of all models is XGBoost, which requires only two features – one process parameter, the agitation setpoint, and one engineering feature, the vessel diameter – to achieve 95%+ characterization of variability and an R^2 value of 0.84. This directly supports the higher-level conclusions that vessel and impeller diameter and impeller power and shear rates are the most important determinants of Performance Indicator 3, since the agitation setpoint directly informs the power and shear rates, while the vessel diameter gives a measure of relative scale. The coefficients and their relative scale can be seen below in Figure 13.

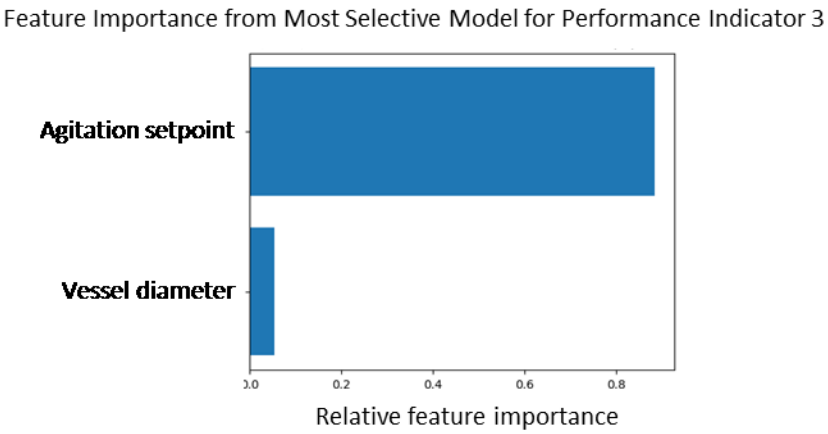

Figure 13. Feature Selection of Most Selective Model for Performance Indicator 3 with Engineering Features. XGBoost achieves an R^2 value of 0.84 by selecting only two features with coefficients greater than the 0.05 threshold. The features selected are the agitation setpoint and vessel diameter, which together provide indirect measures of the power, shear, and relative geometry within the bioreactor.

It is also worth looking at the most selective model among the three top-performers. In this particular analysis that model is Lasso with the default parameters, which selects 10 features. Of these, 40% are engineering features, including that with the largest coefficient: D_i/D_v ratio for impeller 1. As discussed, this is a feature selected by nearly all of the models for this performance indicator, indicating strong predictive power, and is augmented by the inclusion of vessel diameter and max shear rate, both previously discussed, along with the aspect ratio at test volume. This is directly related to the diameter of the vessel, but also incorporates the height of the liquid in the bioreactor (and more specifically, the proportionality of those two values). These engineering features are combined with the process identifier and key process setpoints to achieve an

impressive R^2 value of 0.887 (a near tie with the overall top-performing model for Performance Indicator 3). These coefficients are shown below in Figure 14.

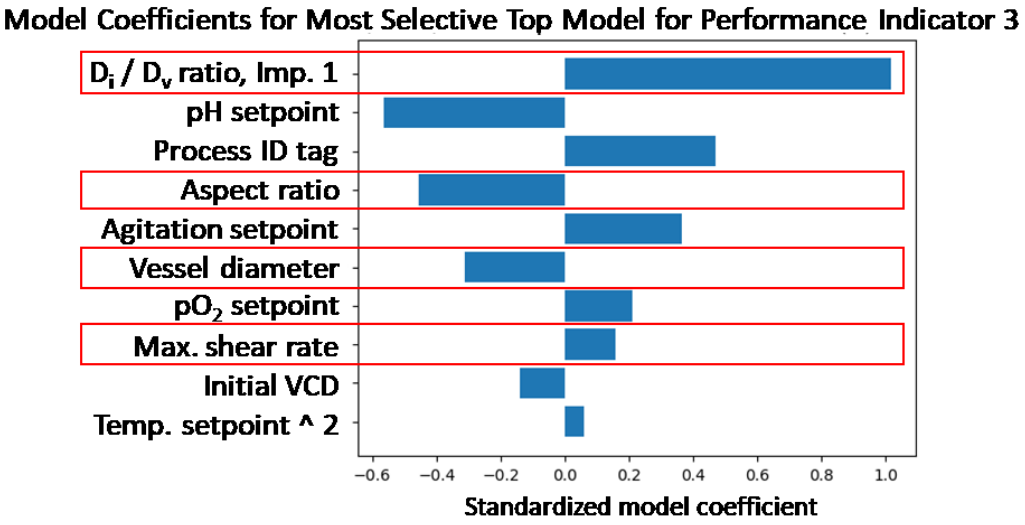

Figure 14. Coefficient Values of Most Selective of Top Three Models Predicting Performance Indicator 3 with Engineering Features. Top models selected by R^2 performance on test set. The variables outlined in red are engineering features which are selected by the Lasso model, constituting 40% of the selected features (out of >50 in the starting data set). These engineering features continue the trend across models of selecting for measures of impeller and bioreactor geometry along with shear, and indicate value gained from the inclusion of engineering features. The features outlined in red are the engineering features selected by the model.

Bayes-Optimized Ridge Regression is the top-performing model for this performance indicator, achieving an R^2 of 0.888 with 13 features selected. The top 10 of those features by magnitude of coefficient are shown below in Figure 15. The features outlined in red are the engineering features selected by the model, which include a direct measure of the total power generated with respect to the volume of the vessel (P/V) and a derivative measure of the same – eddy size, which is calculated as P/V in the context of the fluid density and kinetic velocity – along with a measure of the relative geometry of the impeller and the vessel. This aligns with the previous analyses of features selected across models for this performance indicator.

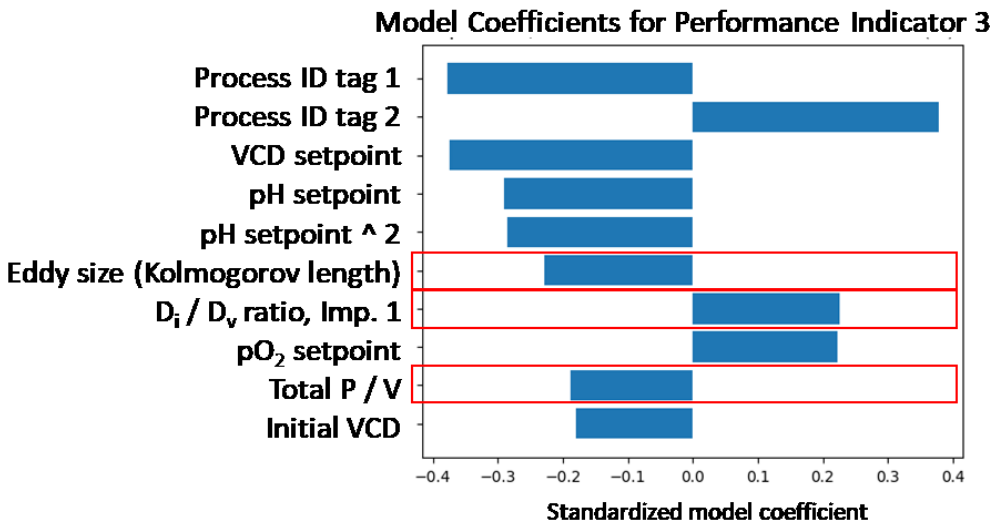

Figure 15. Top 10 Coefficients by Magnitude in Best Model for Performance Indicator 3 with Engineering Features. Best model selected based on R^2 performance on test set. This model, Bayes-Optimized Ridge Regression, selects 13 features in total, with the top 10 shown here. The features outlined in red are the engineering features selected by the model.

In addition to discussing which features come up repeatedly, it is also instructive to consider those that are never selected. For certain direct measures that are not selected, such as liquid height, process working volume, and tip speed, it can be argued that derivatives of each of them were often selected, such as aspect ratio and measures of shear, and thus the information is included in its most valuable form and combination. However, there are other metrics which do not appear in this analysis that are more surprising: Reynolds number, Q_p/V (the specific impeller pumping number), the impeller flow number from which Q_p/V derives, and the blend time. While the information embedded in the Reynolds number is likely not selected because it is sufficiently represented in other features (other combinations of agitation rate and impeller geometry, along with liquid characteristics that remain constant here), the same cannot be stated for impeller flow number, specific impeller pumping number, and blend time. However, the consistency of the impeller setup within a given bioreactor and lack of variability among impeller types may have deemed this information unnecessary, and required simply the geometric ratios that indicate which setup is being considered. It is also possible that the type of flow created by the impeller is representative across the two scales studied, or that it is simply less impactful than the setpoints in concert with the relative geometries, power, and shear forces.

5.1.5 Engineering Feature Paradox

There are two seemingly paradoxical results that arise from this analysis of the impact of engineering features at a general level. First, the impact is variable by response variable and algorithm, and the magnitude of the overall impact is usually rather small on a by-algorithm or by-response variable basis. Second, the top-performing algorithms in the analyses with engineering features – which also tend to be those aggressive on feature selection – select engineering features over process parameter setpoints in many cases, and the selection is often correlated with higher performance.

However, this analysis also reveals that the core of these findings appear to break down most accurately by response variable. Thus, more broadly it can be stated that the impact of engineering features on the whole may vary for each performance indicator related to the bioreactor. In this case, Performance Indicators 1 and 2 are generally adequately predicted by their process parameter setpoints across scales, while Performance Indicator 3 is inadequately predicted by its setpoints and requires further information on the physical and biochemical environment of the bioreactor to inform its predictions. It is interesting to note that the reason for this could be related to the more limited dataset in terms of number of observations, or the fact that Performance Indicator 3 is the only response variable representing an analytical result (a Product Quality Attribute, or PQA), which may be more sensitive to the physical and biochemical drivers described by the calculated engineering features.

It is also possible that these findings are directly related to the level of consistency achieved by Amgen across its scales of manufacturing. In this case, it would indicate that Amgen's bench-scale bioreactors and the parameters at which they are controlled, in concert with their scale-up methodology, create environments that are effectively the same when it comes to measures of process efficiency and productivity, and thus that all of the information required to predict the output is inherent in the setpoints of each scale, and the additional engineering feature information is unnecessary for further characterization.

It is worth noting here that these conclusions may not be robust to different sets of information. For example, if different engineering features were included, if a larger dataset

capturing more variability across scales and process was in place to potentially lend more power to the calculated engineering features in predictions, or if Performance Indicator 3 itself had a richer dataset (meaning more observations), it could change these findings. It is also possible that these changes would only strengthen the findings here, and that these response variables could be representative of the other performance indicators in their category – in particular, with Performance Indicators 1 and 2 representing process efficiency-driven performance indicators and Performance Indicator 3 representing product quality and analytical results (which may be more impacted by the physical and biochemical factors captured by the engineering features).

Thus, it is clear that further study of these relationships is necessary, particularly in a way that can enable a richer, more variable dataset that spans products, processes, and several different scales.

5.2 PROCESS TRANSFERABILITY

The final scenario analyzed in this research is focused on creating an initial test of the possibility for transfer learning using this data and analysis structure. This was done by separating the two processes included in this analysis, training across small and commercial-scale for Process 1 and only small-scale data for Process 2, and then testing on the commercial-scale data for the second process. Thus, the intention is for the models to be able to learn about scale-specific differences from the small-scale and commercial-scale data set of Process 1, and separately learn about the process-specific differences from the small-scale data sets of Process 1 and 2, and be able to apply those learnings in combination to predict at commercial-scale for Process 2.

5.2.1 Performance in Commercial Test Set for New Process

The process transferability scenario is found to fail in this case. With the given data, all models deliver negative R^2 values, indicating that predicting the mean value would have been more accurate. This is true both with and without engineering features included. In fact, the impact of engineering features is negative for Performance Indicators 1 and 3, and makes no difference to Performance Indicator 2.

Thus, it is not possible to show the feasibility of transfer learning among similar processes for use with a brand new process with is scale of data. However, it is in some ways unsurprising that this is not possible, given such a small data set offering extremely limited training data to characterize the process and scale-related sources of variability. We believe that a robust set of multi-process, multi-scale training data could enable this method and be quite powerful in improving cross-scale predictions for new processes. It is also worth recalling that in the baseline scenario assessments, significant improvements in performance arose from the inclusion of data from similar processes; this also supports the hypothesis that a robust data set could enable cross-process learning. Future work could further explore this hypothesis in more detail, with expanded data sets.

5.3 CHALLENGES AND LIMITATIONS

The primary limitation of this analysis lies in the limited amount of data and variability along each axis of interest included in the dataset. There are two main reasons for this limitation: the limited variability in certain biomanufacturing process data and the time-consuming nature of collecting, understanding, and cleaning the historical data sets in concert with the limited timeline of the project.

5.3.1 Limited Variability Inherent in Data Set

The first limitation – limited variability in the data set – is inherent in the nature of biologics manufacturing. In commercial-scale biomanufacturing, the objective is to produce a large amount of a single product in a process that takes place often over multiple weeks, is as controlled, predictable, and replicable as possible. Thus, there are limited runs from which to collect data for analysis – with even blockbuster products generally run only 12-20 times per year – and tight controls and standards to correct and/or abort any nonconformances result in minimal variability in inputs or outputs.[25] Additionally, any irregular inputs or intermediate values related to problems in the manufacturing process will result in terminated runs and incomplete data sets, removing the possibility of correlating final output values with this uncharacteristic variability. In small-scale manufacturing, on the other hand, a small set of prescribed experiments (designed using Design of

Experiments, or DOE, methodology as previously discussed) are used to elucidate relationships between inputs and outputs within set ranges. This data provides a key source of variability from which to learn, but is both restricted to the small-scale environment and limited in quantity, since prior knowledge and experiments and scientist expertise enable expedited processes, while business needs require both speed and efficiency in time and resources. Thus, the data set acquired has two pieces of the puzzle – variability in inputs and outputs, and variability in scales. However, these sources of variability are limited (to two processes and two scales) and do not intersect. That is to say, there is no input/output variability at commercial-scale, nor engineering-feature variability at small-scale. Thus, achieving generalizable learnings that characterize the multiple sources of variability within different contexts becomes challenging.

5.3.2 Limited Size of Data Set and Inclusion of Only Two Scales

Clearly, additional data collection to supplement this variability is desirable, specifically to include more than two processes and scales to assist in elucidating the nature of the process-specific relationships and more generalizable scaling relationships. Two limitations restrict the feasibility of such additional data collection.

The first limitation is related to the inclusion of additional processes. In this case, including additional processes (and thus biological products) would introduce significant additional variability which would be difficult to characterize and separate from the effects of scale on which this project has focused. For example, it would be necessary to understand how different the two products are – in terms of the drug modality, biological characteristics, specifics of the process development, cell culture environments, equipment used, and more, particularly over varying time periods, and determine what data would need to be included to appropriately capture those relationships. In addition, any process included would need to have robust data available not only from small-scale experiments but also commercial-scale manufacturing. These data sets are quite limited due to the length of the processes and scale of manufacturing, and there are not many drugs in Amgen's portfolio with extensive data sets over multiple years of production. Thus, there are few additional products that could be included, each of which would introduce significant sources of additional variability that would be more difficult to understand and characterize. Thus, the

selected single-product, two-process dataset offered the best balance of limited biochemical variability and maximal process variability within the same product.

The second is a practical limitation – that of time. Specifically, this project has been conducted within a six-month timeframe, during which the problem understanding, framing and structuring, interviews, data gathering, data cleaning, data creation in the form of mini-mechanistic bioreactor models and engineering feature calculations, hybrid model creation, scenario creation, and data analysis enabling and embodying this research has been completed. This effort has focused on historical datasets extracted from multiple sources, including many individual e-lab notebooks and other data sources across different time-frames and IT systems. Gathering, cleaning, and connecting the datasets made it necessary to restrict the scope of the data collection portion of this research. Given the previous decision on the processes to be included, this was done by limiting the data collection to two scales – the lab bench and commercial manufacturing.

5.3.3 Matrix Size

As a direct result of the above limitations in the number of observations included, as well as the need to include several key process parameters and interaction terms in addition to the calculated engineering features, the size of the matrix used for this analysis has been subject to "the curse of dimensionality," a core problem in data analysis. A typical rule of thumb for machine learning applications is to have a ratio of 5:1 observations to features, with more observations generally preferred.[26] This analysis has worked to maintain a ratio of nearly 4:1 or higher throughout, which is not ideal but workable, and has focused on regularized regression and machine learning techniques specifically to deal with this matrix size and determine which features are most important and which can be reduced. It is possible that this curse of dimensionality impacts the results – in particular, it may contribute to the higher R^2 values seen in the datasets without engineering features, which universally are found to have a higher ratio of observations to features due to the lack of extra variables. It does not appear that this is a primary driving factor; as discussed in Section 3.3.2, Partial Least Squares regressions are included as a comparison point with reduced dimensionality, and their relative underperformance is interpreted as an indication that the dimensionality is not the cause of decreased performance.

However, it is possible that dimensionality plays some role; as such, it would be interesting for future work to include an analysis of whether including only specific, related subsets of the engineering features and/or increasing or decreasing the observations to maintain an equivalent ratio lead to different results.

Chapter 6: Conclusions and Recommendations

This study has investigated the use of multivariate data analysis, hybrid modeling, and advanced algorithmic techniques to improve the predictive power of models linking process controls to Performance Indicators across scales of production in biomanufacturing cell culture processes. It has sought to investigate the current baseline model and determine if it could be applied to commercial-scale processes, and tested several scenarios to determine if new advanced models could improve predictive power over this baseline. Findings of each of these tests revealed answers to stated questions, while also revealing new questions and hypotheses for further study. These conclusions and topics for further study are discussed in this final chapter.

6.1 Summary of Findings

This study has sought to determine if advanced modeling techniques and hybrid modeling (through the inclusion of supplementary mechanistic-model-based information) could improve predictive power across scales over the current baseline.

Thus, the first step was in establishing and testing the current baseline in this context. This involved using only the data from small-scale experiments in process characterization (and in some sub-scenarios, a small sample of commercial-scale runs from clinical testing). In doing so, it becomes clear that while linear regression analyses performed during process characterization work well on data sets at the same scale – where they are currently used to characterize variability to then apply to expectations of commercial-scale performance, they cannot be applied with any fidelity to predicting performance at commercial-scale, even when small samples of commercial-scale runs are included (as demonstrated by all R^2 values being below 0, and thus worse than predictions of the mean). Thus, with the current limitations in experimental design coverage and data availability, we find that the baseline linear regression approaches are not able to provide cross-scale performance predictions at a higher-fidelity level than the current technique of predicting the mean and mapping variability across scales.

The next step was investigating whether more advanced algorithmic techniques – specifically, regularized regression models such as Lasso, Ridge, and Elastic Net, ensemble of

trees models such as XGBoost, and a simple neural network – could improve predictive performance with the same limited data set. Here, we find that the answer is yes: clear performance improvements are seen in all cases over the baseline regression techniques. However, while these performance improvements are significant, the actual performance of these models is also lacking, with R^2 values for the three response variables falling in the range of -0.044 to 0.007, and thus offering no clear superiority to predictions of the mean. This indicates that within a limited data set of one to two processes and one to two scales of production, more advanced algorithmic techniques cannot adequately discern process-specific and scale-specific relationships between process inputs and outputs, and attempting to extrapolate from small-scale to commercial-scale performance beyond mean values is not possible. Ultimately, the analytical techniques are limited by the quantity and richness of the data, and additional input parameter range variation is needed to characterize these relationships.

The next piece of the analysis focused on investigating hybrid modeling, and the specific impact of the inclusion of calculated engineering features in the data set. For the purposes of enabling these analyses, here, the data set was expanded to include all available data at both scales of production; as in other analyses, this data was then split into training and test sets for fitting and evaluating the models. The availability of data across both scales in concert with the advanced algorithmic techniques improves performance significantly, as expected – with R^2 values of up to 0.91. Thus the focus of these tests was on comparing performance before and after the inclusion of engineering features in the data set. While this analysis required commercial-scale data and thus would not be directly relevant to new processes, this same type of analysis could be used for technology transfer applications, and/or could potentially be performed before commercial production with two or more scales preceding commercial production (such as small-scale models and other intermediate scales). In terms of top-performing models, we find that Bayes optimization for hyperparameter selection is the most successful. In addition, L2 penalty methods – specifically, Bayes-optimized Ridge Regression – perform best with the process-parameter-only data sets (sans engineering features), while L1 penalty methods – in particular, XGBoost and Bayes-optimized Lasso – performed best with the larger data set. Also notable is that the best-performing models always outperform PLS regression, indicating that a different combination of features or reduced-dimensionality problem would not necessarily perform better. With regard to the impact of

engineering features, the findings vary significantly by response variable. Performance Indicator 2, in particular, is well-predicted by its process parameter setpoints (and often has the best performance of all three response variables), and generally does not select for or benefit from any engineering features. Performance Indicator 1 has mixed but relatively insignificant impact from engineering features. Performance Indicator 3, on the other hand, which represents the one analytical product quality result with the largest variance and the smallest available data set, is initially the worst-performing model in the baseline and benefits significantly from the inclusion of engineering features. Models that select for engineering features outperform those that do not, and key patterns emerge with regard to which features are selected: geometric ratios connecting the vessel and impeller geometry and measures of shear stress, along with measure of power, appear to be the key drivers of performance improvements.

These findings indicate that there is significant promise in this hybrid modeling approach, in particular for certain performance indicators similar to Performance Indicator 3, especially those displaying high variability (even within small-scale data sets alone), which are currently inadequately predicted by process setpoints, and/or which suffer from a smaller data set. This is potentially also true for analytical product quality attributes, which may benefit significantly more from additional information on the physical and biochemical environment of the cells in the reactor. Further study will be necessary to test these new hypotheses and extend the analysis to additional performance indicators, particularly product quality attributes and high-variance outputs.

Finally, we endeavored to test whether transfer learning between processes could be used to provide better predictions for a new, but similar process. For this scenario, models were trained on data from Process 1 at both scales and Process 2 at only small-scale, then tested on Process 2 at commercial-scale. In this case, predictive performance is worse than predicting the mean across all response variables. This is similar to the finding regarding the impact of advanced algorithmic techniques in scenarios in which models are trained with extremely limited commercial-scale data: it is determined that there is simply insufficient quantity and richness of data in this case to enable elucidation of input-output relationships across scales and processes. However, it is worth noting that previous analyses – specifically the baseline assessments – do demonstrate that the

supplementation of the data set from a similar process do improve performance significantly; thus, the hypothesis remains that a more robust multi-product multi-scale data set could enable a powerful application of cross-process transfer learning.

6.1.1 New Hypotheses for Future Testing

There are several new hypotheses that arise from this analysis and require further testing. These can be sorted into three categories: 1) that these findings hold with expanded data sets, 2) that the Performance Indicator-specific findings represent patterns that can be extrapolated to similar categories of PIs, and 3) that performance of these models can be improved.

The first category relates to testing these findings against larger and more varied sets of process data. Specifically, the two questions requiring study in this category are whether these findings remain true when a) process data from one or more intermediate scales of production are included, and/or b) process data from one or more additional manufacturing processes are included. We hypothesize that performance improvements and a bigger impact from the inclusion of engineering features would arise with a larger and more robust dataset, capturing more data on both sources of variability under study (process-specific and scale-specific). An additional area of exploration is related to the inclusion of both similar and dissimilar processes (as measured by a key characteristic such as modality), and the number of such processes necessary to fully leverage the benefits (in particular of dissimilar processes).

The second category relates to the response variable-specific findings. Here, the primary question is whether the finding that Performance Indicator 3 (the only Product Quality Attribute or "PQA", that is, output focused on product quality as opposed to process efficiency) benefits the most from the inclusion of engineering features is representative of all PQAs, and that not making use of these features is representative of all Performance Indicators related to process efficiency. We hypothesize that this is not the case, and that the primary reason for the positive impact of engineering features in the case of Performance Indicator 3 is not that it is a PQA, but that it has a smaller dataset with the most variability, and thus cannot be easily characterized by the limited process inputs. Thus, the hypothesis to test here would be that PIs with larger ranges of variability benefit more from engineering features.

The third category relates to opportunities to improve the overall performance of the models studied here. The first opportunity has already been discussed: augmenting the data set with more scales and more processes. The hypothesis here is that the additional data would significantly improve predictive performance. A secondary hypothesis for this scenario is that this more robust data set (particularly with regard to multiple scales) would lead to a higher feature selection rate and more significant impact from the inclusion of engineering features (which would now have more than two sets of values and enable more inherent mathematical scaling relationships to arise). Another opportunity that merits discussion is the use of different subsets of calculated engineering features as opposed to the full set. This would reduce the number of features included in the analysis as well as the multicollinearity of the data set, and thus we hypothesize that clearer patterns and potential performance impacts could be discerned across different methods of regularization and prediction.

6.1.2 Recommendations and Opportunities for Further Development

Based on these findings and the new hypotheses discussed above, there are several opportunities to further develop this analysis. The primary recommendation is to expand the data set by including additional scales of production and additional processes; this will enable a better understanding of both process-specific and scale-specific relationships. In addition, it is recommended that the other hypotheses listed above be tested by investigating new Performance Indicators (both PQAs and process efficiency-related PIs) and subsets of engineering features.

With regard to data set expansion, the recommended long-term approach is to include data from *many* products that have been or are being manufactured, to maximize observations and optimize the matrix size while increasing process-specific variability. This should also include multiple scales of production, such as those used in pre-clinical and clinical manufacturing, pilot labs (throughout scale-up), and small-scale models, to elucidate clearer relationships between calculated engineering features and process outputs, and enable expansion to new equipment, facilities, and scales. Additionally, the data set should contain a ratio of observations to features of at least 5:1. It is worth noting that it would be necessary to add features in the data set identifying modality and other key biological characteristics when additional biological variability is

introduced with the inclusion of more dissimilar processes, but these are likely to be limited in comparison to the increase in observations.

The data set could also be expanded along another dimension, by including continuous or time-series data to better capture the changing environment within the bioreactor. As previously discussed, this could limit the direct application in situations where this data is not available – such as for new processes not yet run at commercial scale. However, these models could still enhance our understanding of the drivers of the final performance indicators under consideration and reveal key time-based patterns. Additionally, for the use cases in which these models are used to enhance and complement first-principles models or for simulation purposes, adding this rich set of information could contribute to far more robust and potentially more accurate models.

Given the challenge and time-consuming nature of collecting the recommended additional data, particularly for small-scale and intermediate-scale runs, we recommend a step-wise approach. This would involve starting first by including more scales for the same processes included here, and then adding in data from similar processes within the same modality, and then adding in full multi-process data sets from other modalities. This will enable the introduction of new sources of variability in a graduated way, enabling more control and potentially more understanding of which pieces are most critical of high-fidelity predictions.

In the long-term, once the data set begins to grow significantly and several processes are included, it will likely make sense to move to a true transfer learning framework, where the models are trained on the very large set of historical process data once, and then additional sets of training data (such as the lab-scale and intermediate-scale data for a new product) are used to update the models before use. As discussed in Section 4.1.7, this architecture is common in image recognition machine learning algorithms, where models will be pre-trained on a very large set of images, and then updated based on a much smaller set of images specifically relevant to the task at hand (such as two breeds of cats before a task focused on discerning between the two).

Another potential long-term approach is the combination of this type of data-driven or hybrid modelling with complex first-principles models of the internal functionality of bioreactors (which focus on the physical, chemical, and biological dynamics and relationships in an attempt to fully simulate the process, and are simultaneously in development). In this approach, key equations

and/or new calculated engineering features from the first-principles models could be incorporated into the machine learning models, and/or machine learning models could be utilized to determine the fitted parameters in first-principles models of bioreactors. This is a potential application under active discussion within Amgen today.[25]

6.2 ORGANIZATIONAL RECOMMENDATIONS

Three additional related recommendations are based on this work: 1) continue to invest in centralized data management and assembling multi-process, multi-scale historical data; 2) establish a central team to act as an Artificial Intelligence / Machine Learning Center of Excellence; and 3) establish clear processes for managing, handing over, and tracking models created for process development and simulation. This section discusses each of these recommendations in the stated order.

6.2.1 Data Centralization

Amgen is a leader amongst its peers in investing in establishing query-able databases that enable various teams to use and analyze common sets of validated data through a data lake. It is critical that this effort, and those to improve accessibility, comprehensiveness, and usability – particularly as relates to easily pulling common fields in cross-process data sets for applications like these – continue. For the long-term vision of this model in use to be realized, it will be necessary to invest in migrating historical process data from the various systems into the same easily-accessible form; this is particularly true for lab-scale data, pilot-scale data, small-scale model data, and other intermediate-scale data. This is the only way that these models will be deployable at-scale in such a way that actionable learnings can arise from multi-process multi-scale multi-site data sets.

6.2.2 Machine Learning Center of Excellence

The second recommendation is to establish a single center of expertise to support all teams within Amgen working on applying artificial intelligence and machine learning techniques to solve problems. In the course of this project, this researcher had the opportunity to interact with many

brilliant scientists and engineers across the company, many of whom had their own models and/or were applying various types of machine learning and other advanced statistical techniques to their own problems. It was helpful to be able to learn from these colleagues, and at times to share learnings and advice back with them. The experience reveals a clear opportunity: employees all over the company are developing models and learning from them, but these learnings tend to stay within each team. If Amgen created a centralized team to help track the efforts of the various teams in this arena, and provide support on best practices and learnings of fellow co-workers, it is believed that all teams could learn, improve, collaborate, and reach their goals more efficiently and effectively.

In concert with the following recommendation, this team could also be responsible for managing a simple database – even an Excel sheet – of the different teams working on machine learning projects, their objectives, and a repository of their models for future reference.

6.2.3 Model Transfer and Management

The final recommendation is focused on managing models created internally to answer a specific question or address a problem. Currently, these models are often created ad-hoc, and lost after the problem is solved that one time by that one employee; or, the employee leaves and completes a haphazard transfer (or none at all) to a colleague. This often leads to not-well-understood Frankenstein models – models added onto but where the initial model is not fully understood and the result is confusing, inelegant, and inefficient – or lost models, along with quite a bit of repeated work as each scientist or engineer builds their own model to solve the same objective over time. This was encountered repeatedly during the course of this project, including as rarely-used repeatedly-handed-off historical analytical scale-up models.

There is a clear opportunity to create and maintain a central repository for analytical models (which could be handled by the Center of Excellence team discussed in Section 6.2.2), with standards for creating adequate documentation and labeling, and for transferring the model to the appropriate person. This will enable all teams to be aware of models previously created or currently in process by other teams, enabling more collaborations, more efficiency, less repeated effort, and a much faster learning cycle.

Simultaneously, it is critically important to begin tracking the engineering features and key bioreactor dimensions, characteristics, and configurations across the manufacturing network. Doing so in a standardized way – as part of a digital equipment footprint and inventory – will ensure all scientists and analysts are working with the same numbers, particularly when working to transfer a process, enabling faster and more standardized analyses across the company while reducing risks and increasing flexibility within the network.

6.3 Business Use Cases and Potential Impact

With further development, we believe that this type of multivariate data analysis leveraging hybrid modeling, machine learning, and in the long-term, transfer learning techniques, can transform the way that Amgen develops new products and processes, and transfers them throughout its global manufacturing network.

6.3.1 Applications of the Model

Several specific applications are envisioned for this type of model. In the short-term, three key uses could bring significant benefits.

1) De-risking scale-up and Process Performance Qualification for new products by improving predictability of new processes at each scale and in each set of equipment after a certain stage of development (leveraging models that have learned from scale-up relationships for similar processes).
2) Increasing the flexibility of the manufacturing network while de-risking technology transfer and pursuant Process Performance Qualification (through both the scale-down step to transfer to the small-scale model at the new facility and the scale-up step to move from that small-scale model to commercial-scale production, along with any intermediate pilot scale in the process). This would involve also taking advantage of the calculated engineering features and multi-site data set, enabling a clearer understanding of site-specific and equipment-specific differences and improving predictability across sites.
3) Identifying potential problem areas for new processes or processes about to be transferred, enabling process scientists to experiment, adapt, and focus efforts on the key areas.

In the longer-term, this modeling approach could be used with other models and analytical tools, such as models developed in other MIT Leaders for Global Operations (LGO) research projects. These could include a predictive model focused on raw material attributes by Emilia Maria Lopez, LGO 2019 and furthered by Zihaui Liu, LGO 2020, and a predictive model focused on leveraging Prior Knowledge Analysis in preparation for process characterization by Or Dan, LGO 2020. In combination, these models could create a comprehensive model across processes, sites, and scales. This model could be used for technology transfer processes, scale-up and scale-down processes, nonconformance investigations, process improvement simulations, or for any other number of prospective or retrospective analyses.

6.3.2 Business Impact

All of the proposed applications of future iterations of this model discussed in Section 6.3.1 would be expected to bring both significant savings and critical strategic value, both minimizing downside and maximizing upside potential.

In terms of savings, successful application of these models could reduce the resource intensity and direct cost of scaling up a process. A large part of this is due to the expected reduction in not-for-human use runs required across multiple scales to characterize a new process and address unexpected variances before the regulated PPQ process. Based on multiple interviews with process scientists and review of documents from the selected processes' scale-up documentation, improved predictability and understanding of process performance across scales could reduce the number of engineering runs necessary by up to 5-10x in some cases. Critically, it could also reduce enormous downside risk: better predictability across scales would reduce the risk of unexpected results, variances, or out-of-specification results during a PPQ – which could cost the company tens of millions of dollars in unusable materials, sunk costs, investigations, re-runs, and lost revenue.

This capability could also provide significant upside in terms of increased manufacturing capacity and faster development timelines (critical in the increasingly competitive biopharmaceutical industry) from fewer manufacturing resources being devoted to process characterization and scale-up, increased flexibility in the manufacturing network from the ability

to transfer processes with fewer resources and less risk, and more resources devoted to bringing new products and more efficiently-produced medicines to market.

In addition, a clearer understanding of the factors that matter the most – in terms of the equipment and engineering characteristics that are most important for consistency, best ways to translate across different manufacturing sites and platforms, the biochemical factors that truly determine key outputs, and more – could provide enormous potential for both cost savings and insights for the development of new molecules and manufacturing processes. It could also make process improvements and iterations less resource-intensive and risky, while improving flexibility in the manufacturing network by making it easier to move processes among different manufacturing platforms, equipment, and sites. Eventually, with sufficient cross-process data, it could enable simulation activities and virtual exploration of different process control operating ranges for a given process.

6.4 CONCLUSION

This research project has identified a clear opportunity to improve predictability of cell culture processes across scales in biologic drug manufacturing, with the potential to impact how new processes are developed, improved, and assessed, and how products move within the manufacturing network. These findings indicate that there is significant promise in this hybrid modeling approach, particularly for selected performance indicators, and more broadly in the use of advanced algorithmic techniques and multivariate data analysis to directly predict cross-scale results. Further study is required to test the several new hypotheses that arose through this work, and several specific ways to expand this analysis in order to test these hypotheses and others in the future have been proposed, all of which will require further investment in expansion of the data set. We look forward to seeing how this type of analysis develops in the industry in the coming years, and are grateful for the opportunity to contribute to pushing it forward.

References

1. 7 biopharma trends to watch in 2019 | BioPharma Dive. https://www.biopharmadive.com/news/7-biopharma-trends-to-watch-in-2019/546011/.
2. Biotechnology | Definition of Biotechnology by Lexico. *Lexico Dictionaries | English* https://www.lexico.com/en/definition/biotechnology.
3. Timeline | An Introduction to Biotechnology. https://www.biotechnology.amgen.com/timeline.html.
4. Evens, R. & Kaitin, K. The Evolution Of Biotechnology And Its Impact On Health Care. *Health Aff. (Millwood)* **34**, 210–219 (2015).
5. IBISWorld - Biotechnology in the US Market Size 2005-2025. https://www.ibisworld.com/industry-statistics/market-size/biotechnology-united-states/.
6. Biotechnology Market Size, Growth Forecast is Projected to be Around US$ 795.7 billion by 2026 - MarketWatch. https://www.marketwatch.com/press-release/biotechnology-market-size-growth-forecast-is-projected-to-be-around-us-7957-billion-by-2026-2019-12-30.
7. Research, C. for B. E. and. What Are 'Biologics' Questions and Answers. *FDA* (2019).
8. Biologics: Definition, Side Effects, Uses & Drug List. https://www.medicinenet.com/biologics_biologic_drug_class/article.htm.
9. Blackstone, E. A. & Joseph, P. F. The Economics of Biosimilars. *Am. Health Drug Benefits* **6**, 469–478 (2013).
10. How the U.S. Compares to Europe on Biosimilar Approvals and Products In the Pipeline. *Rothwell Figg Blog* https://www.biosimilarsip.com/2019/05/07/how-the-u-s-compares-to-europe-on-biosimilar-approvals-and-products-in-the-pipeline-4/ (2019).

11. The Shape of Drugs to Come. *Amgen Science* https://www.amgenscience.com/features/the-shape-of-drugs-to-come/.

12. Strube, J., Grote, F., Josch, J. P. & Ditz, R. Process Development and Design of Downstream Processes. *Chem. Ing. Tech.* **83**, 1044–1065 (2011).

13. Clapp, K. P., Castan, A. & Lindskog, E. K. Upstream Processing Equipment. in *Biopharmaceutical Processing* 457–476 (Elsevier, 2018). doi:10.1016/B978-0-08-100623-8.00024-4.

14. Bioreactor. *Wikipedia* (2020).

15. Mercier, S. M., Diepenbroek, B., Dalm, M. C. F., Wijffels, R. H. & Streefland, M. Multivariate data analysis as a PAT tool for early bioprocess development data. *J. Biotechnol.* **167**, 262–270 (2013).

16. *Chemical engineering in the pharmaceutical industry: R&D to manufacturing*. (Wiley, 2011).

17. Scale, Standardize, or Normalize with Scikit-Learn - Towards Data Science. https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02.

18. How to use Data Scaling Improve Deep Learning Model Stability and Performance. https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/.

19. (Tutorial) Regularization: Ridge, Lasso and Elastic Net - DataCamp. https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net.

20. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825−2830 (2011).

21. XGBoost Algorithm: Long May She Reign! - Towards Data Science. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.

22. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785–794 (ACM Press, 2016). doi:10.1145/2939672.2939785.

23. Chollet, F. & others. *Keras*. (2015).

24. A Gentle Introduction to Transfer Learning for Deep Learning. https://machinelearningmastery.com/transfer-learning-for-deep-learning/.

25. Cell-Culture Advances Test Bioreactor Performance Models - Bioprocess Development Forum. http://www.processdevelopmentforum.com/articles/cell-culture-advances-test-bioreactor-performance-models/.

26. Theodoridis, S. & Koutroumbas, K. *Pattern recognition*. (Academic Press, 2009).