Improving Prior Knowledge Assessment in Process Characterization

by

Or Dan

B.Sc. Industrial Engineering and Management, Tel Aviv University, Israel (2014)

Submitted to the MIT Sloan School of Management and the Operations Research Center in Partial Fulfillment of the Requirements for the Degrees of

Master of Business Administration
and
Master of Science in Operations Research

In conjunction with the Leaders for Global Operations Program at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2020

© 2020 Or Dan. All rights reserved.
The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created,

Signature of Author ____________________________
MIT Sloan School of Management, Operations Research Center
May 8, 2020

Certified by ____________________________
Retsef Levi
J. Spencer Standish (1945) Professor of Operations Management
MIT Sloan School of Management
Thesis Supervisor

Certified by ____________________________
John N. Tsitsiklis
Clarence J Lebel Professor
Department of Electrical Engineering and Computer Science
Thesis Supervisor

Approved by ____________________________
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

Approved by ____________________________
Georgia Perakis
William F. Pounds Professor of Management Science, co-director, Operations Research Center
This page intentionally left blank.
Improving Prior Knowledge Assessment in Process Characterization

by

Or Dan

Submitted to the MIT Sloan School of Management and the Operations Research Center on May 8th, 2020 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Operations Research

Abstract

A critical aspect of biologics manufacturing is creating a safe, reliable and consistent manufacturing process. The manufacturing process design includes process characterization (PC) experiments to demonstrate process robustness and provide data necessary for planning, risk mitigation, development of the control strategy, and successful execution of process validation.

Performing PC experiments is resource intensive, both human and capital, so leveraging prior knowledge from previous experiments is essential. Until now, using data from past experiments relied on a centralized static document called Prior Knowledge Assessment (PKA). The PKA aggregates the results of many statistical models that were created during past PC studies.

Using the PKA provides insight, but leaves a lot of room for subjective decision making around questions, such as: How should products be grouped together? and What operating parameters are more important? The PKA also lacks uncertainty quantification for statistical significance.

In this thesis, we aggregated data from past PC experiments across multiple molecules, and developed a machine learning framework to holistically analyze cross-product data from process characterization DOE studies.

The model developed through this project provides interpretable predictions of sensitivity of Performance Indicators to Process Parameters variation. The model enables, for the first time, to assess and quantify the impact of parameters on indicators, even if they were not tested originally for a specific molecule. A novel user interface was created in order to bring the framework to life and create a “one-stop shop” for a scientist to interact with the model. This work improves process characterization decision quality. Potential benefits of this approach would be to increase speed and agility in process development and reduce number of future experiments.

Thesis Supervisor: Retsef Levi
Title: J. Spencer Standish (1945) Professor of Operations Management
MIT Sloan School of Management

Thesis Supervisor: John N. Tsitsiklis
Title: Clarence J Lebel Professor
Department of Electrical Engineering and Computer Science
Acknowledgements

I would like to thank my MIT academic advisors, Retsef Levi and John N. Tsitsiklis, for their advice, mentorship and for the technical insights provided during the internship and while writing the thesis.

Secondly, I want to thank Amgen Inc. and my supervisor Roger Hart for tailoring the opportunity of working on such an interesting project and for the support of this thesis. I would like to thank specifically Paul Gramlich, Catherine Grimm and Alex Jacobitz, for their continuous guidance, encouragement and candid feedback to ensure a successful project. Their patience and support have been critical to completing this project.

I am grateful for the support structure that Amgen Inc. established for the MIT LGO interns. Dollie Grajczak was instrumental in making our experience unique from start to end. I would also like to thank Aine Hanly, our LGO executive sponsor at Amgen Inc., who was involved with all of our projects. The LGO alumni community Amgen Inc. provided a fantastic support network and I am grateful for it.

My fellow February-Start Amgen Inc. interns, Aaron Baskerville-Bridges, Nicole Oliver and Zoe Wolszon provided continuous feedback, insights and invaluable support. I feel blessed to be a part of the larger LGO family.

Lastly, I would like to thank my family and friends. I am humbled to be at MIT and the encouragement and support from them are the very things that make this journey possible.

“If I have seen further it is by standing on the shoulders of giants” – Isaac Newton
The author wishes to acknowledge the Leaders for Global Operations Program for its support for this work.
# Table of Contents

Abstract ......................................................................................................................... 3  
Acknowledgements ........................................................................................................ 4  
List of Figures .................................................................................................................. 8  
List of Tables .................................................................................................................. 9  
1 Introduction ................................................................................................................. 10  
   1.1 Project Motivation .................................................................................................. 10  
   1.2 Project objective ..................................................................................................... 11  
   1.2.1 Significance of Study ......................................................................................... 12  
   1.3 Thesis Overview .................................................................................................... 13  
2 Background ................................................................................................................. 14  
   2.1 Biopharmaceutical Industry .................................................................................... 14  
   2.2 AMGEN ................................................................................................................. 16  
   2.3 Biologics Manufacturing ....................................................................................... 16  
   2.3.1 Upstream and downstream process .................................................................. 20  
   2.3.2 Cation Exchange Chromatography (CEX) ...................................................... 22  
   2.4 Molecular descriptors .......................................................................................... 24  
   2.5 Process Development ........................................................................................... 25  
   2.5.1 Process Characterization ................................................................................. 26  
   2.5.2 Prior Knowledge Assessment (PKA) ................................................................ 27  
3 Literature Review ........................................................................................................ 31  
4 Approach ..................................................................................................................... 33  
   4.1 Framework ............................................................................................................ 33  
   4.2 Data and Tool Architecture .................................................................................. 33  
   4.2.1 Data collection and preparation ....................................................................... 34  
   4.2.2 Process Parameters (input data) ....................................................................... 35  
   4.2.3 Molecular Descriptors (input data) .................................................................. 36  
   4.2.4 Performance Indicators (output data) .............................................................. 36  
   4.2.5 Data Sparsity ..................................................................................................... 37  
   4.3 Approaches assessed ............................................................................................ 38  
   4.3.1 Models assessed ............................................................................................... 39  
   4.3.2 Models configuration and design ....................................................................... 40  
   4.3.3 Metrics used in model selection ...................................................................... 41  
5 Results ........................................................................................................................ 43  
   5.1 Model performance in predicting performance indicators ..................................... 43  
   5.2 Contribution of molecular descriptors ................................................................. 47  
6 User Interface and implementation ............................................................................ 52  
   6.1 Creating a middle layer for data transformation .................................................. 52  
   6.2 Choosing Python programing language .............................................................. 53  
   6.3 Choosing Spotfire analytics platform .................................................................... 53  
   6.4 The Analysis Tool ................................................................................................. 53
Recommendations and Contributions

7.1 Future work: expand research

7.1.1 Expand to other unit operations

7.1.2 Model Unit Operations dependency

7.1.3 BiTE® molecule evaluations

7.2 PD and AMGEN Recommendations

7.2.1 Ensure data governance and Model Governance

7.2.2 Datalake paradox – improving centralized data projects

7.2.3 Develop skills for maintenance and upgrade

7.3 Major Contributions

7.3.1 Multi-Product / Multi Parameter / Multi Indicator Model

7.3.2 Developed a user interface

References

Glossary
List of Figures

Figure 1. General Scheme of upstream and downstream processes. [7] ......................................... 19

Figure 2. Classic biologics manufacturing process (Amgen Inc., 2018c) . ........................................... 20

Figure 3. Chromatographic Separation – molecules move in different speeds. a) The molecule that has greater attraction for the mobile phase will exit the column faster b) the other component. .......................................................................................................................... 23

Figure 4. The isoelectric point boundary for protein charge ................................................................ 24

Figure 5. Current PKA – Separate models results are aggregated into a document ......................... 28

Figure 6. PKA table for one performance indicator............................................................................ 29

Figure 7. PIR – Process Impact Rating ............................................................................................... 30

Figure 8. X marks that a Parameter or an Indicator value data was recorded for a specific molecule during the Process Characterization study. ................................................................. 38

Figure 9. General machine learning approach .................................................................................... 39

Figure 10. Diagram of cross-validation with k=5 .............................................................................. 41

Figure 11. New state marked with GREEN arrows: machine learning replaces the individual models that were created separately and introduces dynamic user interface instead of static document. ......................................................................................................................... 52

Figure 12. A snapshot of the analysis tool (values and names masked) ............................................ 54

Figure 13. This part displays the data sparsity table, as shown in section 3.2.5 ............................... 55
List of Tables

Table 1. Process Parameters  ................................................................. 35
Table 2. Performance Indicators ........................................................... 36
Table 3. Machine Learning models evaluated ....................................... 40
Table 4. Displays the Performance Indicator Metrics, sorted by RMSE / Range Size. 43
Table 5. Displays the Performance Indicator Metrics, sorted by Winning Model. 44
Table 6. Displays the Performance Indicator Metrics, sorted by Total number of Observations.46
Table 7. Displays Indicators simulation results, sorted by indicator name ................................. 48
Table 8. Displays the correlation coefficients matrix across molecules ......................... 51
Table 9. Detailed area by area explanation of the tool snapshot ......................... 54
1 Introduction

With the cost of developing a prescription drug at more than $2 billion, and a period of 10-plus years spent in development phase, pharmaceuticals companies are constantly looking for opportunities to minimize cost and accelerate speed of development. Innovative approaches of applying machine learning to existing data sets that have never been combined before, could offer a way to achieve competitive advantage and support this audacious effort.

During development of a new biologic drug, manufacturing in initial phases is done at a small scale. While scaling-up the manufacturing of a new molecule from small scale to commercial scale, manufactures, such as Amgen, need to demonstrate that the manufacturing process remains safe, consistent and reliable.

The goal of Process characterization (PC) experiments is to demonstrate process robustness and provide data necessary for planning, risk mitigation, development of control strategy, and successful execution of process validation. Amgen has accumulated a large body of data through the characterization of many therapeutic protein processes. This organizational knowledge is gathered for each manufacturing stage. It is used to develop more efficient PC control strategies and better establish the company’s process understanding.

In a typical experiment, a group of scientists change one or more Process Parameters (input), such as temperature, or pH level and measure Performance indicators (output), such as Yield. Performing PC experiments is both human and capital resource intensive, so leveraging prior knowledge from previous experiments is essential.

1.1 Project Motivation
At its core, process characterization is about creating a consistent, reliable and safe manufacturing process. Until now, using data from past experiments data is based on a centralized static document called Prior Knowledge Assessment (PKA). The PKA aggregates the results of many statistical models that were created as part of past PC studies. Using the PKA provides insight, but leaves room for subjective decision making.

If an operating parameter (OP) has a positive and a negative impact on different products, which product is more important when evaluating a new molecule candidate? Or in other words, how should products be grouped together? what operating parameters (OP) are more important? It also lacks quantification for statistical significance. Maintaining the PKA updated with newest DOE information has also proven to be a very time-consuming and organizationally challenging process.

On the one hand, maintaining safety and meeting regulation requirements increase the pressure to produce a lot of data in order to demonstrate a deep understanding and control of manufacturing processes. On the other hand, as the pace of biotechnology quickens, shortening time to market and resource saving by reducing the number of experiments would be beneficial. As Amgen expects to have a larger pipeline of molecules in the coming years, an opportunity arises to leverage historical data for a predictive analytics approach in process characterization.

1.2 Project objective

The objective of this project is to develop a framework to holistically analyze cross product data from process characterization studies. The framework should help to inform how to design the experiments for a new candidate molecule, as part of the design and assessment of the (control) parameters of the to be developed manufacturing process.
The input to the model would be results from experiments done on prior molecules, combined with molecular qualities called molecular descriptors. The output would be the predicted value for the desired performance indicators for a new molecule with little or no experiment data of its own.

The hypothesis is that combining the experimental data and descriptors, and using linear and tree based machine learning methods, would allow to create one centralized model that will improve process characterization decision quality, by quantifying the impact of parameters on indicators, even if they were not tested originally for a specific molecule.

The model developed through this project should provide interpretable predictions of sensitivity of Performance Indicators to Process Parameters variation. The framework shall enable the development of an integrated business process to support Prior Knowledge Assessment. We aspire that the benefits of this approach would be to increase speed and agility in process development and reduce number of future experiments.

1.2.1 Significance of Study

The combination of Process Characterization studies data with molecular descriptor data based only on machine learning yielded a multi-product (molecule), multi-parameter and multi indicator model. No specific algorithm has proven to be the best for all Indicators, but complex models such as Regression Trees and XGBoost were generally more successful.

Using the metric of RMSE / Indicator Range Size, out of 22 Performance Indicators, top 17 have a score lower than or equal to 20%, top 8 have a score lower than or equal to 10%. And the bottom 5 indicators had 71 data points or less.
The model enables **for the first time** to assess and quantify the impact of parameters on indicators, even if they were not tested originally for a specific molecule.

This study applies machine learning to create a novel application in biologics manufacturing. Conservatively, we present proof-of-concept rollout of our methodology, with acknowledgement that additional vetting would be required in order to utilize the framework in process characterization efforts. Optimistically, we foresee benefits that extend beyond our initial goals. The framework we create will be modular, which means we can effectively expand it to other stages and processes to further its scope in the future.

More generally, we hope to shed light on the possibilities and opportunities of machine learning in biotechnology.

### 1.3 Thesis Overview

This thesis consists of seven chapters. Chapter 2 provides necessary background information of biologics manufacturing and process characterization. In chapter 3 we provide literature review sources that shed more light on biologics manufacturing and reference specific technical references for machine learning choices we made.

In chapter 4 we outline the approach to the framework we developed. We discuss how the data was obtained and transformed. We then detail the design choices in refining our algorithms, and metrics used. In chapter 5 we discuss our results, including model performance and insights. Chapter 6 details the design of the front-end application created to provide access to the model to a scientist. Chapter 7 lists recommendations and contributions, we complete the thesis with discussions of successes, shortcomings, and recommendations for future work.
2 Background

This chapter provides a broad background of the Biopharmaceutical industry, biologics manufacturing and Process Development. In each section we provide a more in-depth overview of essential areas for this work such as CEX, Process Characterization and molecular descriptors.

2.1 Biopharmaceutical Industry

The Biopharmaceutical industry is comprised of companies engaged in researching, developing, manufacturing and distributing drugs for human or veterinary use. New drugs have an enormous positive influence on global health, prosperity and economic productivity by saving lives, increasing life spans, reducing suffering, preventing surgeries and shortening hospital stays. Advances in medicine have eliminated deadly diseases and have brought other life-threatening conditions under control. Drug therapy is now an integral part of nearly every facet of healthcare, and new breakthroughs promise to revolutionize the treatment of non-communicable diseases. (Department of Commerce, USA)

Biologics (biotech drugs, biological drugs, biopharmaceuticals) include a wide range of products such as vaccines, therapeutic proteins, blood and blood components, tissues, etc. In contrast to chemically synthesized drugs, which have a well-defined structure and can be thoroughly verified, biologics are derived from living material (human, animal, microorganism or plant) and are vastly larger and more complex in structure. Biologic medicines revolutionize the treatment of cancer and autoimmune disorders and are critical to the future of the industry. (Department of Commerce, USA)

Traditional drugs are based on small molecules that typically have a lower molecular weight (<1000 Da) and are delivered in pill form. Their size allows them to be processed into tables or
capsules. If the pill can dissolve in the gastrointestinal tract, then the active substance is absorbed into the bloodstream and from there it can reach almost any desired destination in the human body due to the tiny size. Small molecules can sometimes penetrate cell membranes.

In contrast, biologics consist of large protein molecules with a typical molecular weight of 150,000 Da, more or less, that have a desired therapeutic effect. They bind to specific receptors on or in the cells that are associated with the process of the disease. The ability to bind large molecules selectively to a small subset of abnormal cells makes them highly effective and causes fewer side effects. Large molecules are administered by injection or infusion in order to avoid the destruction of the protein structure in the stomach and intestines. Unlike small molecules that are chemically synthesized, large molecule drugs are derived from living systems. Bacteria, yeast, or Chinese Hamster Ovary (CHO) cells are engineered with a specific DNA that yields the desired protein. The cells produce the protein as they are grown. The protein is then harvested, purified and formulated so it can be used as treatment for patients.

The pharmaceutical industry is one of the most heavily regulated sectors in the world. Drugs are evaluated for safety, efficacy, and manufacturing quality; misleading product claims and illicit inducements to choose a particular drug are prohibited. Prices are regulated in many countries through their respective healthcare and insurance systems. While product success in the U.S. market is largely determined by open competition based on quality, safety and price, internationally, companies face a patchwork of uneven regulations, protectionist policies and price controls. These obstacles are increasingly being instituted in both developing and developed countries. Regulatory complexity and efforts to contain accelerating health costs are key challenges the U.S. industry needs to overcome. (Department of Commerce, USA)
As of 2019, major players in the pharma and biopharma space include Pfizer, Merck, Johnson & Johnson, Roche, Sanofi, Novartis, AbbVie, AstraZeneca, Gilead Sciences, and Amgen.

2.2 AMGEN

AMGen (Applied Molecular Genetics Inc.) was established in Thousand Oaks, California, in 1980, as the brainchild of venture capitalists William K. Bowes and associates. In 1993 the company raised a $40 million IPO and officially changed its name to Amgen Inc. (Amgen Inc., 2018a). The company grew to become a world-class pioneering biotechnology company. With presence in approximately 100 countries and regions worldwide, as of December 2019, Amgen Inc.’s market capitalization exceeds $140 billion (Yahoo Finance, December 2019)

Amgen is committed to unlocking the potential of biology for patients suffering from serious illnesses by discovering, developing, manufacturing and delivering innovative human therapeutics. Amgen’s belief—and the core of their strategy—is that innovative, highly differentiated medicines that provide large clinical benefits in addressing serious diseases are medicines that will not only help patients, but also will help reduce the social and economic burden of disease in society today. (Amgen, Inc., 2018)

Amgen Inc. is a leader in the manufacturing of biologics, which represents 76 % of its product portfolio (Amgen Inc., 2018b). Amgen focuses on six therapeutic areas: cardiovascular disease, oncology, bone health, neuroscience, nephrology and inflammation. The medicines typically address diseases for which there are limited treatment options, or they are medicines that provide a viable option to what is otherwise available. (Amgen, Inc., 2018)

2.3 Biologics Manufacturing
This section provides a high-level explanation of biologics manufacturing. It covers upstream and downstream process in general, and then introduces the Cation Exchange Chromatography (CEX) purification step in order to contextualize the environment for this work.

Biologics are produced in living cells and are inherently complex due to naturally occurring molecular variations. The manufacturing process requires highly specialized knowledge at the intersection of biology, manufacturing and advanced process control. Extensive process and product characterization are required to transform laboratory-scale processes into reproducible commercial manufacturing processes. Conditions must be precisely monitored in order to produce quality products consistently. At a high level, the manufacturing process consists of the following four main steps (See 1-1):

1. **Host cell selection**: Producing the master cell line containing the gene that makes the desired biologic; Chinese hamster ovary (CHO) cells are typically used for these cells.

2. **Cell culture (Upstream)**: Using defined culture media to grow large numbers of cells that produce the protein in big tanks called Bioreactors.

3. **Purification (Downstream)**: The recovery and the purification of biosynthetic products from the previous stage in order to isolate product protein.

4. **Formulation - Fill-Finish-Package**: Preparing the protein for use by patients.

First, an optimal cell which produces the target protein in high concentration is engineered. Cells are cryopreserved in vials or cell bags (~2 - 5 mL). Upstream manufacturing is initiated with the revival of the cryopreserved cells. The cells are thawed into small T-flasks, shake flasks, or spinner flasks and expanded in increasing numbers and sizes of flasks to achieve inoculation.
of a seed bioreactor. Throughout the expansion process, the cells are kept in optimal conditions (temperature, pH, nutrients) for continued growth. Following the inoculum expansion and seed bioreactor steps, the cells are inoculated into the production bioreactor. During the production bioreactor step, the therapeutic protein is expressed by the cells. Following this step, in downstream step, the desired protein is isolated from the cells and the growth media. Various filtering technologies are used to isolate and purify the proteins based on their size, molecular weight, and electrical charge. Lastly, the purified protein is typically formulated with an excipient to produce a sterile solution that can be injected or infused. The final steps are to fill vials or syringes with individual doses of the finished drug and to label the vials or syringes, package them, and make them available to physicians and patients.
Figure 1. General Scheme of upstream and downstream processes. [7]

The described procedure is essentially a batch process. Once a batch progresses through one stage, the equipment is either cleaned or discarded, in the case of Single Used Systems, and prepared for the next use. At every step of the upstream process, maintaining the specific environment that cells require in order to thrive is critical. Even subtle changes can affect the cells and alter the proteins they produce. Due to the high sensitivity, strict controls are needed to ensure the quality and consistency of the final product in accordance with Good Manufacturing Practice (Figure 1-2). Amgen Inc. carefully monitors process variables such as temperature, pH,
nutrient concentration, and oxygen levels throughout the manufacturing process. Frequent tests are also run to guard against contamination from bacteria, yeast, and other microorganisms.

2.3.1 Upstream and downstream process

As stated above, the two main steps in biologics manufacturing are called upstream and downstream. During host cell selection step, production clones have been selected and
characterized. The outcome of the step is a working cell bank of frozen vials containing the cells to be used.

To begin the **upstream** process for a production run, a vial would be thawed. From this vial, the manufacturer will propagate and expand the cells required for production. Cells are expanded in test tubes, shake flasks, or disposable bags, and then into progressively larger bioreactors, from a few liters in volume, up to as much as 20,000 liters.

This expansion stage could consist of five or six steps in sequence. Each of these steps may take hours or days depending on the growth rate of the culture. A general rule of thumb for mammalian cells is that the next reactor in the train should be 5 to 10 times larger in volume than the current one. The emphasis in the expansion stage is on growing sufficient numbers of cells, while that in the production stage is total product yield and quality.

Important factors during the upstream process are the sourcing and sterilization of input materials used to create the media, the preparation and storage of the solutions used, and cleaning protocols for equipment between batches.

The core of the upstream process is the production bioreactor. Controlling these factors is critical to optimize the productivity of the bioreactor. The manufacturer strives to maintain quality while maximizing the total yield of product for a given volume over a period of time. In other words, maximize the titers of the products while minimizing time length and cost of production in terms of raw materials, energy supply and labor.

The **downstream** process is essentially a series of multiple steps to recover the desired biologic from the upstream process and meet purity and quality requirements. Since a single
purification step is not enough to achieve the required purity level, different separation processes are used.

Proteins expressed in CHO cells are secreted to the culture medium and separated from cells by centrifuge or filtration operations; this is the case for processes analyzed within this thesis. Cell disruption techniques are used if the desired product is intra cellular, so that the product is released. Cell disruption can be carried out via mechanical or non-mechanical methods. Mechanical methods could induce shear force to break cell walls. Non-mechanical methods to break down cell walls could include physical methods, such as using steel beads as an abrasive, chemical methods, such as using a detergent, or the use of enzymes. A combination of these methods can lead to a higher release of the product and a more effective result.

Other separation processes rely on leveraging differences between the target product and other molecules in the bioreactor. Size differences could be in captured by separating materials using filtration methods, or it could be carried out by centrifugation. Hydrophobic interaction chromatography allows to separate proteins based on differences in their hydrophobicity. Affinity chromatography separates molecules based on differences in their affinity for a target ligand attached to the chromatography resin, and ion exchange chromatography separates molecules based on differences in their charge.

2.3.2 Cation Exchange Chromatography (CEX)

Chromatography refers to a separation where molecules are distributed between two phases. The first is a stationary phase, often a chromatography resin. The second is a mobile phase, which in the case of protein separation is a solvent, such as water. Molecules that are more attracted to the stationary phase move more slowly through the system than those that are more
attracted to the mobile phase. For commercial purifications, chromatography is always carried out as column chromatography due to scale considerations.

![Figure 3. Chromatographic Separation – molecules move in different speeds. a) The molecule that has greater attraction for the mobile phase will exit the column faster b) the other component.](image)

In a common chromatographic operation, a volume of sample is injected into the column. Eluent is then pumped through the column, causing molecules to be separated based on their relative affinity for the stationary resin and the eluent. Molecules will elute from the column at different times and after different volumes of eluent have passed through the column. This is captured in a chromatogram, which is a plot of the concentration exiting the column versus time.

Cation-exchange chromatography (CEX) is an ion exchange chromatography used when the molecule of interest is positively charged. Proteins have amino acids with acidic and basic side chains. Depending on the acidity level (pH) of the solution surrounding the biologic, it can be positively charged, negatively charged, or neutral.

The isoelectric point, or pI, is the pH at which the number of protonated and deprotonated groups is equal, and the protein has no net charge. If the pH is greater than the pI, a protein will have a net negative charge, and if the pH is less than the pI, the protein will have a net positive charge.
Figure 4. The isoelectric point boundary for protein charge

Because a protein’s pI is determined by its primary amino acid sequence and can thus be calculated, a buffer can then be chosen that ensures a known net charge for a protein of interest. Proteins with different pI values will have varying degrees of charge at a given pH and so different proteins will bind to the resin with different strengths, facilitating their separation through the column.

2.4 Molecular descriptors

Molecular descriptors are formal mathematical representations of a molecule obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure: “The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment.” [9]
A single molecule could be described in many different ways. Thousands of numerical descriptors could be computed for any given chemical. Many of these descriptors are very closely related to each other and might capture the same information at times. Thus, the selection of relevant descriptors is a well-known problem.

In general, a lot of experience is required in order to select the appropriate descriptors for a model development for a specific use-case. The nature of the chemical structure being considered has to be taken into account. A set of descriptors may efficiently encode the chemical information perfectly for the small molecules, but the same set of descriptors may not be able to encode the required features for polymers, protein structures, and inorganic molecules. Thus, the selection process of appropriate descriptors requires additional knowledge beyond the knowledge required to calculate them. [3]

Since some descriptors depend on the charge of the protein, and the pH level determines the surface charge of the protein. It is crucial to use pH-dependent descriptors to enhance their relevance to a pH-dependent environment.

2.5 Process Development

In the context of bio manufacturing, Process Development is the discipline of developing the upstream and downstream manufacturing processes for a given molecule. Process Development includes the identification of the required steps in the process, and their corresponding parameters values that are needed to produce a specified molecule. When bringing a new drug to market, drug developers must meet regulatory requirements to prove that their drug and processes are safe, efficacious, reliable and of high quality. In order to ensure reliability and quality of the product, the manufacturing process must be controlled and well understood.
Therefore, one regulatory requirement for biopharmaceuticals is that process robustness be demonstrated and scientifically justified by the manufacturer.

Process Development organizations are often engaged early on in the development stage to design and scale a manufacturing process suitable to supporting clinical trials and subsequent commercial supply. After the commercial manufacturing process has been transferred and validated, changes to the existing process must undergo thorough evaluations, and often regulatory reviews, to ensure those changes do not impact the product integrity.

2.5.1 Process Characterization

Initially, biologics are manufactured at small quantities using a small scale manufacturing process. While scaling up the manufacturing of a new molecule from small scale to commercial scale, Amgen needs to demonstrate the manufacturing process continues to be safe, consistent and reliable. One part of the overall process development is called process characterization. During this process, studies and experiments are conducted and data is being gathered. In an experiment, a group of scientists changes one or more Process Parameters (input), such as temperature, or pH level and measure Performance indicators (output), such as Yield. Although some tests are done once (“one-offs”), usually a series of experiments are executed under one design of experiments (“DOE”). Using multifactorial test design allows Amgen to achieve statistically significant results, while exploring multiple parameters combinations.

Depending on the process, manufacturing capabilities and control environment, parameters can only be controlled to a limit. The goals of process characterization studies are to identify process parameters that impact product indicators that usually represent quality and yield. Those studies are used to justify manufacturing operating ranges and acceptance criteria. Interactions
between process parameters and critical quality attributes should be identified as well, all to ensure that the output of the manufacturing process delivers the required molecule with reproducible results.

2.5.2 Prior Knowledge Assessment (PKA)

As mentioned above, Process characterization (PC) experiments are performed to demonstrate process robustness and provide data necessary for planning, risk mitigation, and successful execution of process validation.

Amgen has accumulated a large body of information through the characterization of many therapeutic protein processes. This organizational knowledge is gathered for each manufacturing stage. It is used to develop more efficient PC control strategies and better establish the company’s process understanding.

Until now, using data from past experiments data was facilitated by a centralized document called Prior Knowledge Assessment (PKA). Data granularity is exchanged for a more holistic view. A linear regression model was computed for each DOE, and the results of the models were aggregated into the PKA. The PKA would then inform about the overall impact of process parameters on performance indicators.
PKA aggregates the results of many statistical models that were computed for every series of experiments, the table in Figure 6 below captures whether a certain parameter has a positive, negative or neutral (insignificant) correlation with a certain performance indicator.
**Figure 6. PKA table for one performance indicator**

**Figure 6** shows a PKA illustration. For example, the top, second from the left cell, colored gold, with a minus signs means that Operating Parameter 2 ("OP2") has negative, significant (Cell color Gold = PIR 3) impact on product A. for the performance indicator described by the entire table.

Such a table has to be computed for each performance indicator, for each manufacturing stage. **Figure 7** displays how the impact is normalized to a grade between 1 and 4 according to the scheme below:

**Process impact rating (PIR):** An operating parameter rating that is a function of the sensitivity process performance to perturbations of the operating parameter.
Normal operating range (NOR): The operating range for a parameter that is found in manufacturing batch records (typically based on equipment tolerances and capabilities rather than process capability); the range is generally two-sided (± around a set point). “1×, 2×, . . . NOR” refers to how an operating parameter is set relative to the NOR for a given study. For example, if the pH set point and NOR is 7.5 ± 0.2, then a study testing the pH range of 7.1–7.9 would be a “2×” NOR study.

![Figure 1: PIR scoring table; scores go from 4 (highest impact) to 1 (lowest impact).](image)

*Figure 7. PIR – Process Impact Rating*
3 Literature Review

The overall goal of this work is to leverage the experiments historical data and molecular descriptors data, and apply machine learning to create a predictive model that can improve process characterization decision quality. The four components of this work: CEX, process characterization, molecular descriptors, and machine learning are all domains with an extensive research activity.

The challenge of working with molecular descriptors is well documented. Researchers used descriptors generated by MOE, and electron-density-derived descriptors in conjunction with genetic algorithm and partially least squared regression to reduce number of descriptors, and developed a model to predict CEX retention time. They tested 22 proteins of different kinds, CHO was not included in the sample [1].

The notion of using appropriate large molecule descriptors that are pH dependent in CEX setting has been suggested as a way of improving predictive power [2].

Researchers have developed a prediction model of retention time using three molecular descriptors: molecular weight (MW), charge, and surface hydrophobicity (SH). The research was conducted on corn protein and included three data sets at the size of 497, 132 and 41 proteins. Using leave one out cross validation and regression tress, it highlighted the advantage of using regression tress to handle non linearity [3].

Creating and maintaining prior knowledge assessment documents, using aggregation of process characterization models results and normalizing them to a shared scale can inform future process characterization efforts [4].
The complexity of biologics manufacturing and its inherent variability is the motivation to develop a predictable, controlled process [5].

This thesis introduces several extensions to previous research done in this domain. First, we have developed a machine learning model for a small subset of molecules. Second, we have done so for varying environment conditions, whereas most other researches use constant conditions between different molecules. Third, we have developed a model for a variety of indicators simultaneously. Lastly, none of the previous works has been implemented as an analytics tool that a scientist uses to assist in decision making.
4 Approach

This chapter describes the input data and output data in more detail, the architecture that transforms the input into the output, and models assessed.

4.1 Framework

The general framework used could be summarized using the following flow chart:

This framework addresses the following aspects:

- **Updateability** – new studies are conducted continuously. We designed the framework to enable incorporation of new data with minimal effort, rather than a manual “snapshot” of currently available data.

- **Flexibility** – Not only new data entries can be incorporated easily; new features could be introduced as well.

- **Traceability** – The system uses data that is verified by scientists. Since the system manipulates data and should inform decisions all steps are completely traceable and visible.

- **User Interaction** – User interaction with the model is key to allow scientists to explore results and improve process characterization decision quality.

4.2 Data and Tool Architecture

In this section we will detail the input data and output data, discuss their compositions and elaborating on how they the data was gathered.
4.2.1 Data collection and preparation

Process Characterization studies have been conducted in Amgen in the last few decades. Our approach was to build a model using the original raw data from those experiments, rather than relying on the outputs of the old statistical models as described in the Prior Knowledge Assessment section. This required us to identify the scope and depth of our data collection.

Cation Exchange Chromatography (CEX) purification stage was chosen since it offered relatively large quantities of data compared to other stages. In addition, CEX changes throughout the last decade or so were minor compared to changes and improvements in other manufacturing stages. The rather stable state of CEX enabled us to more realistically assume similar external conditions in studies throughout the years.

PC studies and their attached data are stored in lab notebooks which were converted to electronic lab notebooks (“ELN”) in the early 2000s. Throughout the years, the data was recorded in spreadsheets in multiple ELN systems that replaced one another. Data collection, filtering and standardization was conducted over a period of several months.

While most relevant data was available in spreadsheets, some data had to be manually transcribed from scanned files. High stakeholder engagement was required to determine the location, relevance, and nuances of each dataset. This manual data collection process was effort-intensive, requiring meetings with project owners and scientists across the organization. Eventually, we acquired PC experiments data for 16 molecules, with 954 lines of data. The data standardization process included multiple steps:
• **Filtering** - leaving data to contain only series of experiments. Experiments that were ran only once (“one-offs”) were not included. Additionally, filtering out some unrelated contextual metadata fields related to the experiment logistics.

• **Renaming** – renaming fields to a common name in order to merge data from different studies together.

• **Conversion** – Converting units between different studies.

• **Adding environment data** – During studies, parameters that their values were not changed were not recorded in the experiment spreadsheet, but only in the study report. Those values had to be added as constants to the study spreadsheet.

• **Merging** – All studies were merged to one large matrix. The matrix was joined with molecular feature data.

4.2.2 **Process Parameters (input data)**

For each of the Process Characterization studies, only a subset of the Process Parameters was changed and recorded in its spreadsheet:

<table>
<thead>
<tr>
<th>Process Parameter Name</th>
<th>Measurement Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Buffer pH</td>
<td>pH</td>
</tr>
<tr>
<td>2 Elution Buffer Conductivity</td>
<td>mS/cm</td>
</tr>
<tr>
<td>3 Elution Buffer Molarity</td>
<td>mM</td>
</tr>
<tr>
<td>4 Elution Buffer pH</td>
<td>pH</td>
</tr>
<tr>
<td>5 Gradient Slope</td>
<td>mM/CV</td>
</tr>
<tr>
<td>6 Linear Velocity</td>
<td>cm/hr</td>
</tr>
<tr>
<td>7 Load Conductivity</td>
<td>mS/cm</td>
</tr>
</tbody>
</table>
4.2.3 Molecular Descriptors (input data)

Molecular descriptors are mathematical representations of different molecular attributes. The data set of molecular descriptors was generated using an industry standard software called “molecular operating environment” (MOE). As stated in the Chromatography section, CEX is heavily dependent on pH levels. Therefore, each descriptor was generated in different pH levels so it can be linked accordingly to the changing pH conditions in the input data. [2]

Initially, 245 different descriptors were generated. pH levels were rounded to the first decimal place. For efficiency, during the import of the input data, molecular descriptors that had identical values between all molecules and all pH levels were declared as redundant and dropped from the data set. The final list of molecular descriptors before machine learning algorithms were applied included 78 descriptors.

4.2.4 Performance Indicators (output data)

For each of the Process Characterization studies, only a subset of the Performance Indicators, which represent quality measurements, was recorded in the study spreadsheet:

<table>
<thead>
<tr>
<th>Performance Indicator Name</th>
<th>Total number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Performance Indicator 1</td>
<td>446</td>
</tr>
<tr>
<td>2  Performance Indicator 2</td>
<td>543</td>
</tr>
<tr>
<td>3  Performance Indicator 3</td>
<td>492</td>
</tr>
<tr>
<td></td>
<td>Performance Indicator 4</td>
</tr>
<tr>
<td>---</td>
<td>--------------------------</td>
</tr>
<tr>
<td>5</td>
<td>Performance Indicator 5</td>
</tr>
<tr>
<td>6</td>
<td>Performance Indicator 6</td>
</tr>
<tr>
<td>7</td>
<td>Performance Indicator 7</td>
</tr>
<tr>
<td>8</td>
<td>Performance Indicator 8</td>
</tr>
<tr>
<td>9</td>
<td>Performance Indicator 9</td>
</tr>
<tr>
<td>10</td>
<td>Performance Indicator 10</td>
</tr>
<tr>
<td>11</td>
<td>Performance Indicator 11</td>
</tr>
<tr>
<td>12</td>
<td>Performance Indicator 12</td>
</tr>
<tr>
<td>13</td>
<td>Performance Indicator 13</td>
</tr>
<tr>
<td>14</td>
<td>Performance Indicator 14</td>
</tr>
<tr>
<td>15</td>
<td>Performance Indicator 15</td>
</tr>
<tr>
<td>16</td>
<td>Performance Indicator 16</td>
</tr>
<tr>
<td>17</td>
<td>Performance Indicator 17</td>
</tr>
<tr>
<td>18</td>
<td>Performance Indicator 18</td>
</tr>
<tr>
<td>19</td>
<td>Performance Indicator 19</td>
</tr>
<tr>
<td>20</td>
<td>Performance Indicator 20</td>
</tr>
<tr>
<td>21</td>
<td>Performance Indicator 21</td>
</tr>
<tr>
<td>22</td>
<td>Performance Indicator 22</td>
</tr>
</tbody>
</table>

### 4.2.5 Data Sparsity

As can be seen previously in this section and in Figure 8, the Process Characterization studies data is fairly sparse and quite diverse. Different molecule programs studied different Process Parameters and different Performance Indicators throughout the years.
Figure 8. X marks that a Parameter or an Indicator value data was recorded for a specific molecule during the Process Characterization study.

The sparsity created a challenge for us. Dropping records was not an option due to the already small volume of data. Leaving values as blanks/nulls would make using numerical methods infeasible. In contrary to other use cases, missing values could not be simply replaced with zeros due to the scientific value they represent.

A few different ways of imputing missing values were discussed with the team (K-Nearest neighbors, mean, median, mode, constant), and it was decided that imputing the mean value dynamically for each model would be the only alternative the team would be comfortable with. This was due to the fact it was deemed the closest conservative approximation of the actual conditions.

### 4.3 Approaches assessed

This section is dedicated to outline the details of the machine learning model design. We present the predictive models implemented, the way they were calibrated, and the quality metrics which enabled the best model selection. The general approach can be seen in Figure 10.
The inputs to the model are the parameters conditions in previous experiments on different molecules, combined with molecular descriptors. Those features are used to produce the output of the model, the predicted value of a performance indicator.

4.3.1 Models assessed

The approach taken to model the impact of Process Parameters on Performance Indicators is supervised machine learning. Because the prediction values, as well as all input values were numerical, we utilized regression estimators in our machine learning pipeline. The following models were trained:
Table 3. Machine Learning models evaluated

<table>
<thead>
<tr>
<th>Model Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
</tr>
<tr>
<td>Ridge</td>
</tr>
<tr>
<td>Elastic Net</td>
</tr>
<tr>
<td>Stochastic Gradient Descent (SGD) regularized loss linear model</td>
</tr>
<tr>
<td>Linear Support Vector Machine (SVM)</td>
</tr>
<tr>
<td>Partial least squares regression (PLS)</td>
</tr>
<tr>
<td>Decision / ID trees</td>
</tr>
<tr>
<td>Extreme gradient boosting (XGBoost)</td>
</tr>
</tbody>
</table>

It was important to select different types of models (regression, trees) to allow diversity of approaches to be evaluated and potentially be put to use.

4.3.2 Models configuration and design

In this part, we will elaborate on cross-validation and hyper-parameter optimization.

Cross-validation:

It is the common practice to have training and testing datasets in machine learning. The best performing model on the training set is evaluated by looking at the quality of its prediction on the testing set. An out-of-sample score could then be calculated and evaluated. When data is abundant, the train set could be further split up to a train and validation set to allow assessment of the prediction model before it is evaluated on the “external” test set.

When data is scarce, cross-validation is the standard to be used. For each model, we performed k-fold validation ($k = 10$), evaluating model performance ten times across 90/10 partitions of the data.
Hyper-parameter optimization:

The predictive algorithms we evaluate have different parameters controlling their behavior called hyper-parameters. Those parameters could control for example, the rate of learning, number of iterations, tree-depth and penalty functions. Optimizing the values of hyper-parameters is challenging due to the higher-dimensional search space.

For this project, a Bayesian search method was utilized [8]. This method searches the possible space computationally more efficiently than grid search or random search, and yields similar levels of performance as random search. Bayesian hyper-parameter optimization balances between the method’s design to explore places with high variance and the goal to seek the lowest mean. The models for this work used 10 iterations of Bayesian search, and all model hyper-parameters were chosen this way, through cross-validation (with \( k = 10 \)).

4.3.3 Metrics used in model selection
As outlined during this section, after a variety of models were trained for each performance indicator, a final model had to be chosen. We tracked both $R^2$ and RMSE scores for each iteration and calculated an average score based on the cross validation process. The model that yielded the lowest average RMSE was chosen to be the winning model.

In Equations 1-4, $k$ represents the number of cross-validation folds, $n$ represents the number of samples per fold, $y$ represents the true target output, $\bar{y}$ represents the average of the true target output, and $\hat{f}$ represents the predicted output.

**Equation 1 $R^2$**

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n} (\hat{f}_i - y_i)^2}{\sum_{i=1}^{n} (\bar{y} - y_i)^2}$$

**Equation 2 Average $R^2$**

$$Average\ Model\ R^2 \equiv \frac{1}{k} \sum_{j=1}^{k} R^2_j$$

**Equation 3 RMSE – Root Mean Squared Error**

$$RMSE \equiv \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i - y_i)^2}$$

**Equation 4 Average RMSE**

$$Average\ Model\ RMSE \equiv \frac{1}{k} \sum_{j=1}^{k} RMSE_j$$

RMSE was the measure of choice, it reflects accuracy data about the error in original performance indicator units. RMSE was preferred over mean absolute error (MAE) because we wanted to penalize larger errors between predictions and actual data, and it is also better understood by the scientists who interact with the model. The $R^2$ metric on occasion would yield extremely negative values ($R^2 \ll 0$) with some cross-validation sets, which would skew the model comparison when averaged across sets.
5 Results

This section describes the primary results of this work, including developing the framework into a full functional tool for a Process Development scientist to interact with.

5.1 Model performance in predicting performance indicators

As described in the previous section, based on a RMSE score, a winning model was chosen for each performance indicator. The range size for each performance indicator was calculated as the difference between its maximal and minimal value in the entire data set, to contextualize the size of RMSE. RMSE / Range Size does not take into account the actual distribution of the indicator but still provides a rough estimate of performance. The lower it is, the better.

The table below displays the Performance Indicator name, winning model, RMSE, RMSE / range size and total number of observations. Sorted by RMSE / range size.

Table 4. Displays the Performance Indicator Metrics, sorted by RMSE / Range Size.

<table>
<thead>
<tr>
<th>Performance Indicator Name</th>
<th>Winning Model</th>
<th>RMSE</th>
<th>RMSE / Range Size</th>
<th>Total number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Indicator 3</td>
<td>Regression Tree</td>
<td>0.741</td>
<td>2%</td>
<td>492</td>
</tr>
<tr>
<td>Performance Indicator 2</td>
<td>XGBoost</td>
<td>0.845</td>
<td>2%</td>
<td>543</td>
</tr>
<tr>
<td>Performance Indicator 7</td>
<td>Regression Tree</td>
<td>0.035</td>
<td>3%</td>
<td>339</td>
</tr>
<tr>
<td>Performance Indicator 1</td>
<td>XGBoost</td>
<td>0.763</td>
<td>3%</td>
<td>446</td>
</tr>
<tr>
<td>Performance Indicator 5</td>
<td>XGBoost</td>
<td>3.003</td>
<td>5%</td>
<td>879</td>
</tr>
<tr>
<td>Performance Indicator 9</td>
<td>XGBoost</td>
<td>0.239</td>
<td>7%</td>
<td>157</td>
</tr>
<tr>
<td>Performance Indicator 18</td>
<td>Regression Tree</td>
<td>0.442</td>
<td>7%</td>
<td>55</td>
</tr>
<tr>
<td>Performance Indicator 6</td>
<td>Regression Tree</td>
<td>0.225</td>
<td>10%</td>
<td>541</td>
</tr>
<tr>
<td>Performance Indicator 4</td>
<td>Regression Tree</td>
<td>0.247</td>
<td>10%</td>
<td>811</td>
</tr>
<tr>
<td>Performance Indicator 14</td>
<td>Regression Tree</td>
<td>0.059</td>
<td>12%</td>
<td>71</td>
</tr>
</tbody>
</table>
Besides Indicator 18 that has a smaller ratio than Indicators 6 and 4, there is a clear split where the smaller ratios belong to indicators with a significantly higher number of observations. The result that more data drives better performance in machine learning should have no shortage of support in the field. But it still provides a concrete, quantifiable demonstration for this general statement in a Process Characterization setting.

Elastic Net won for 2 indicators, XGBoost won for 8 and Regression trees won for 12 indicators. Elastic Net appears at the bottom of the table 4 but other than that there is no clear dominant model between XGBoost and Regression trees, they are spread out across the table.

Table 5. Displays the Performance Indicator Metrics, sorted by Winning Model.

<table>
<thead>
<tr>
<th>Performance Indicator Name</th>
<th>Winning Model</th>
<th>RMSE</th>
<th>RMSE / Range Size</th>
<th>Total number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Indicator 17</td>
<td>Regression Tree</td>
<td>0.141</td>
<td>16%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 11</td>
<td>Regression Tree</td>
<td>0.399</td>
<td>17%</td>
<td>160</td>
</tr>
<tr>
<td>Performance Indicator 20</td>
<td>XGBoost</td>
<td>0.725</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>Performance Indicator 19</td>
<td>XGBoost</td>
<td>1.203</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>Performance Indicator 8</td>
<td>XGBoost</td>
<td>0.483</td>
<td>18%</td>
<td>86</td>
</tr>
<tr>
<td>Performance Indicator 21</td>
<td>XGBoost</td>
<td>1.214</td>
<td>18%</td>
<td>90</td>
</tr>
<tr>
<td>Performance Indicator 22</td>
<td>Regression Tree</td>
<td>0.340</td>
<td>20%</td>
<td>82</td>
</tr>
<tr>
<td>Performance Indicator 15</td>
<td>Regression Tree</td>
<td>0.272</td>
<td>25%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 13</td>
<td>Elastic Net</td>
<td>0.170</td>
<td>28%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 10</td>
<td>Elastic Net</td>
<td>0.282</td>
<td>28%</td>
<td>41</td>
</tr>
<tr>
<td>Performance Indicator 16</td>
<td>Regression Tree</td>
<td>0.213</td>
<td>35%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 12</td>
<td>Regression Tree</td>
<td>0.044</td>
<td>44%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator</td>
<td>Model</td>
<td>Score</td>
<td>Improvement</td>
<td>Value</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td>--------</td>
<td>-------------</td>
<td>-------</td>
</tr>
<tr>
<td>10</td>
<td>Elastic Net</td>
<td>0.282</td>
<td>28%</td>
<td>41</td>
</tr>
<tr>
<td>13</td>
<td>Elastic Net</td>
<td>0.170</td>
<td>28%</td>
<td>71</td>
</tr>
<tr>
<td>11</td>
<td>Regression Tree</td>
<td>0.399</td>
<td>17%</td>
<td>160</td>
</tr>
<tr>
<td>12</td>
<td>Regression Tree</td>
<td>0.044</td>
<td>44%</td>
<td>71</td>
</tr>
<tr>
<td>14</td>
<td>Regression Tree</td>
<td>0.059</td>
<td>12%</td>
<td>71</td>
</tr>
<tr>
<td>15</td>
<td>Regression Tree</td>
<td>0.272</td>
<td>25%</td>
<td>71</td>
</tr>
<tr>
<td>16</td>
<td>Regression Tree</td>
<td>0.213</td>
<td>35%</td>
<td>71</td>
</tr>
<tr>
<td>17</td>
<td>Regression Tree</td>
<td>0.141</td>
<td>16%</td>
<td>71</td>
</tr>
<tr>
<td>18</td>
<td>Regression Tree</td>
<td>0.442</td>
<td>7%</td>
<td>55</td>
</tr>
<tr>
<td>22</td>
<td>Regression Tree</td>
<td>0.340</td>
<td>20%</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>Regression Tree</td>
<td>0.741</td>
<td>2%</td>
<td>492</td>
</tr>
<tr>
<td>4</td>
<td>Regression Tree</td>
<td>0.247</td>
<td>10%</td>
<td>811</td>
</tr>
<tr>
<td>6</td>
<td>Regression Tree</td>
<td>0.225</td>
<td>10%</td>
<td>541</td>
</tr>
<tr>
<td>7</td>
<td>Regression Tree</td>
<td>0.035</td>
<td>3%</td>
<td>339</td>
</tr>
<tr>
<td>1</td>
<td>XGBoost</td>
<td>0.763</td>
<td>3%</td>
<td>446</td>
</tr>
<tr>
<td>19</td>
<td>XGBoost</td>
<td>1.203</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>XGBoost</td>
<td>0.845</td>
<td>2%</td>
<td>543</td>
</tr>
<tr>
<td>20</td>
<td>XGBoost</td>
<td>0.725</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>21</td>
<td>XGBoost</td>
<td>1.214</td>
<td>18%</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>XGBoost</td>
<td>3.003</td>
<td>5%</td>
<td>879</td>
</tr>
<tr>
<td>8</td>
<td>XGBoost</td>
<td>0.483</td>
<td>18%</td>
<td>86</td>
</tr>
<tr>
<td>9</td>
<td>XGBoost</td>
<td>0.239</td>
<td>7%</td>
<td>157</td>
</tr>
</tbody>
</table>

There are a few interesting results in this table. The first is that using multiple models was helpful in improving performance. There was no one model that won for all indicators. The
winning models were Elastic Net, Regression Tree and XGBoost. On the other hand, Lasso, Ridge, Stochastic Gradient Descent (SGD), Linear Support Vector Machine (SVM) and Partial least squares regression (PLS) models did not win even one indicator.

The second is that although Elastic Net won for two indicators with low volume of data, the more complicated models of Regression trees and XGBoost won in 20/22 indicators, even for indicators with relatively low data (<100). However, the latter models are not currently used in traditional statistical analysis for Process Characterization studies. First because each study has a separate model based on low volume of data (~30-50 data points), and secondly because the expected analysis and standard operating procedure dictate using a simple linear regression analysis.

Table 6 Displays the Performance Indicator Metrics, sorted by Total number of Observations.

<table>
<thead>
<tr>
<th>Performance Indicator Name</th>
<th>Winning Model</th>
<th>RMSE</th>
<th>RMSE / Range Size</th>
<th>Total number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Indicator 5</td>
<td>XGBoost</td>
<td>3.003</td>
<td>5%</td>
<td>879</td>
</tr>
<tr>
<td>Performance Indicator 4</td>
<td>Regression Tree</td>
<td>0.247</td>
<td>10%</td>
<td>811</td>
</tr>
<tr>
<td>Performance Indicator 2</td>
<td>XGBoost</td>
<td>0.845</td>
<td>2%</td>
<td>543</td>
</tr>
<tr>
<td>Performance Indicator 6</td>
<td>Regression Tree</td>
<td>0.225</td>
<td>10%</td>
<td>541</td>
</tr>
<tr>
<td>Performance Indicator 3</td>
<td>Regression Tree</td>
<td>0.741</td>
<td>2%</td>
<td>492</td>
</tr>
<tr>
<td>Performance Indicator 1</td>
<td>XGBoost</td>
<td>0.763</td>
<td>3%</td>
<td>446</td>
</tr>
<tr>
<td>Performance Indicator 7</td>
<td>Regression Tree</td>
<td>0.035</td>
<td>3%</td>
<td>339</td>
</tr>
<tr>
<td>Performance Indicator 11</td>
<td>Regression Tree</td>
<td>0.399</td>
<td>17%</td>
<td>160</td>
</tr>
<tr>
<td>Performance Indicator 9</td>
<td>XGBoost</td>
<td>0.239</td>
<td>7%</td>
<td>157</td>
</tr>
<tr>
<td>Performance Indicator 19</td>
<td>XGBoost</td>
<td>1.203</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>Performance Indicator 20</td>
<td>XGBoost</td>
<td>0.725</td>
<td>17%</td>
<td>90</td>
</tr>
<tr>
<td>Performance Indicator 21</td>
<td>XGBoost</td>
<td>1.214</td>
<td>18%</td>
<td>90</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>Performance Indicator 8</td>
<td>XGBoost</td>
<td>0.483</td>
<td>18%</td>
<td>86</td>
</tr>
<tr>
<td>Performance Indicator 22</td>
<td>Regression Tree</td>
<td>0.340</td>
<td>20%</td>
<td>82</td>
</tr>
<tr>
<td>Performance Indicator 13</td>
<td>Elastic Net</td>
<td>0.170</td>
<td>28%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 12</td>
<td>Regression Tree</td>
<td>0.044</td>
<td>44%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 14</td>
<td>Regression Tree</td>
<td>0.059</td>
<td>12%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 15</td>
<td>Regression Tree</td>
<td>0.272</td>
<td>25%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 16</td>
<td>Regression Tree</td>
<td>0.213</td>
<td>35%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 17</td>
<td>Regression Tree</td>
<td>0.141</td>
<td>16%</td>
<td>71</td>
</tr>
<tr>
<td>Performance Indicator 18</td>
<td>Regression Tree</td>
<td>0.442</td>
<td>7%</td>
<td>55</td>
</tr>
<tr>
<td>Performance Indicator 10</td>
<td>Elastic Net</td>
<td>0.282</td>
<td>28%</td>
<td>41</td>
</tr>
</tbody>
</table>

This table illustrates there is no dominant model between XGBoost and Regression trees when it comes to dataset size, as both win large and small indicators datasets. It also shows that although data set size improves performance in general as discussed before, we can observe that indicators 3, 1 and 7 have a lower RMSE/Range ratio than indicators 5 and 4 despite indicators 5 and 4 having at least 60% more data.

### 5.2 Contribution of molecular descriptors

Molecular descriptors were the data component that enabled us to combine the data from multiple PC studies, from different molecules, into one model. As stated in the background about molecular descriptor in Section 2.4, choosing molecular descriptors is a well-known challenge. While many models in the field of bioprocess development are mechanistic and based on first-principles, our approach was to develop a data-driven predictive framework. In this approach the
algorithm itself chooses the relevant features depending on the data, algorithm and hyperparameters. That selection is likely to change as new data is introduced to the model.

It is not straightforward to evaluate what is the contribution of molecular descriptors across the framework, since the models are not linear. The contribution can be different for different indicators under different experimental conditions. In an effort to estimate some of the difference driven by molecular descriptors we ran the following simulation:

1. Generated 10,000 random records of random experiments conditions, all within the value range of the original process parameters. The number 10,000 comes from an internal practice of simulation during DOE analysis; results were not significantly different when ran with 1,000 random records.
2. Predicted the outcome for each performance indicator, for all molecule programs.
3. Calculated the average predicted value for each molecule-indicator pair.
4. Calculated the average and standard deviation for every indicator across all molecules.
5. Calculated the coefficient of variation (CV) for every indicator, CV is the standard deviation divided by the mean and is used as a measure of variation magnitude.

Table 7. Displays Indicators simulation results, sorted by indicator name

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Average across Molecules</th>
<th>Standard Deviation</th>
<th>Coefficient of Variation (CV)</th>
<th>Total number of Observations</th>
<th>Indicator Range Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Indicator 1</td>
<td>25.42</td>
<td>5.56</td>
<td>0.22</td>
<td>446</td>
<td>24.94</td>
</tr>
<tr>
<td>Performance Indicator 2</td>
<td>63.68</td>
<td>11.22</td>
<td>0.18</td>
<td>543</td>
<td>41.70</td>
</tr>
<tr>
<td>Performance Indicator 3</td>
<td>15.59</td>
<td>10.53</td>
<td>0.68</td>
<td>492</td>
<td>29.70</td>
</tr>
<tr>
<td>Performance Indicator 4</td>
<td>0.80</td>
<td>0.50</td>
<td>0.62</td>
<td>811</td>
<td>2.40</td>
</tr>
<tr>
<td>Performance Indicator 5</td>
<td>91.98</td>
<td>3.69</td>
<td>0.04</td>
<td>879</td>
<td>58.94</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>Performance Indicator 6</td>
<td>99.22</td>
<td>0.46</td>
<td>0.00</td>
<td>541</td>
<td>2.16</td>
</tr>
<tr>
<td>Performance Indicator 7</td>
<td>0.40</td>
<td>0.49</td>
<td>1.21</td>
<td>339</td>
<td>1.27</td>
</tr>
<tr>
<td>Performance Indicator 8</td>
<td>96.16</td>
<td>0.04</td>
<td>0.00</td>
<td>86</td>
<td>2.69</td>
</tr>
<tr>
<td>Performance Indicator 9</td>
<td>2.06</td>
<td>0.01</td>
<td>0.01</td>
<td>157</td>
<td>3.21</td>
</tr>
<tr>
<td>Performance Indicator 10</td>
<td>98.12</td>
<td>0.00</td>
<td>0.00</td>
<td>41</td>
<td>1.00</td>
</tr>
<tr>
<td>Performance Indicator 11</td>
<td>-42.24</td>
<td>269.90</td>
<td>-6.39</td>
<td>160</td>
<td>2.40</td>
</tr>
<tr>
<td>Performance Indicator 12</td>
<td>0.94</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>0.10</td>
</tr>
<tr>
<td>Performance Indicator 13</td>
<td>29.31</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>0.60</td>
</tr>
<tr>
<td>Performance Indicator 14</td>
<td>2.30</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>0.50</td>
</tr>
<tr>
<td>Performance Indicator 15</td>
<td>69.08</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>1.10</td>
</tr>
<tr>
<td>Performance Indicator 16</td>
<td>-4.30</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>0.60</td>
</tr>
<tr>
<td>Performance Indicator 17</td>
<td>97.36</td>
<td>0.00</td>
<td>0.00</td>
<td>71</td>
<td>0.90</td>
</tr>
<tr>
<td>Performance Indicator 18</td>
<td>13.14</td>
<td>0.00</td>
<td>0.00</td>
<td>55</td>
<td>6.40</td>
</tr>
<tr>
<td>Performance Indicator 19</td>
<td>21.00</td>
<td>0.08</td>
<td>0.00</td>
<td>90</td>
<td>7.10</td>
</tr>
<tr>
<td>Performance Indicator 20</td>
<td>5.22</td>
<td>0.24</td>
<td>0.05</td>
<td>90</td>
<td>4.20</td>
</tr>
<tr>
<td>Performance Indicator 21</td>
<td>72.67</td>
<td>0.16</td>
<td>0.00</td>
<td>90</td>
<td>6.60</td>
</tr>
<tr>
<td>Performance Indicator 22</td>
<td>192.43</td>
<td>388.20</td>
<td>2.02</td>
<td>82</td>
<td>1.70</td>
</tr>
</tbody>
</table>

The results confirm the claims made by Amgen scientists throughout the project, and also demonstrate some of the framework limitations:

1. Most of the indicators with low volume (<100) of data have CV = 0. We do not interpret that they are all the same but that the data was insufficient for the winning model to
include and express molecule dependency. Most of those indicators also have a very small range.

2. Indicator 11 (160 data points) and Indicator 16 (71 data points) averaged a negative value which is not scientifically feasible. Although the model was trained with only non-negative values, negative values were predicted due to the combination of low data volume, and the breath of generating 10,000 different experiment conditions that are broader than the original study conditions. It demonstrates the weakness of applying machine learning without incorporating domain specific knowledge.

3. Indicator 6 with high volume of data (541 data points) and very low CV attracted our attention. But the low CV is explained by the relative large Indicator values compared to the range size of 2.16. This can be seen by comparing to another similar high volume data Indicator 4 (811 data points), which has a similar range size (2.4 vs 2.16) and standard deviation (0.5 vs 0.46). It is easy to see that the low CV is attributed to the fact that Indicator 6 average value is 99.22 compared to Indicator 4 with average value of 0.8.

4. Until now, decisions about how to draw conclusion by grouping molecules together were based on scientific understanding and experience. The results per molecule (not shown at the table above) display clearly how some molecules are similar to others per indicator, or a group of indicators with similar predictions that are significantly different than other molecules. Table 8 displays the correlation coefficients across molecules. The ability to quantitatively observe those similarities would be useful when designing a new study or evaluating performance for an indicator that was not tested for a specific molecule.
5. The result that molecules are different from one another is obvious, but it is still interesting to witness that even in 10,000 random different conditions the average for many of the indicators has considerable variation.
6 User Interface and implementation

One of the goals of this project was to provide a Process Development scientist a dynamic tool to assist in decision making. In this section we will detail some of the technical components and the rationale for choosing them for this project. The components could be replaced by other components or tools for similar scenarios.

Figure 11. New state marked with GREEN arrows: machine learning replaces the individual models that were created separately and introduces dynamic user interface instead of static document.

6.1 Creating a middle layer for data transformation

Once the Process Characterization studies spreadsheet were collected, they had to be saved to a shared folder as they could not be queried directly from the electronic lab notebook. A master spreadsheet points to all the spreadsheets and contains all encoded instructions for data transformation. It also stores additional meta data about the study, such as the ELN ID and
program name. The master sheet was shared with the team. It allowed to keep track of the progress, and enabled collaboration in the data gathering and transformation efforts.

6.2 Choosing Python programming language

Once it became apparent that a heavy manipulation would be required and some of the data might be unstructured, the decision was made to go with Python. Python offers more manipulation options with other operating system components. It was also the recommended language at Amgen at the time.

The Python program reads the master sheet and then automatically imports all the required data while following the transformation instructions. A great deal of effort was spent to make the program as dynamic as possible, new studies could be simply added to the master sheet and then be automatically incorporated into the model. The models used in this project were implemented using sklearn Python package.

6.3 Choosing Spotfire analytics platform

With the models trained in python, they were many options of exporting and deploying web applications or analytics applications for the user. Spotfire analytics is the premier platform used by Amgen Process development organization and serves a variety of analytics needs on the top of the enterprise data lake. The scientists are familiar with it and comfortable with it, so we decided to adopt it in order to deliver a solution which would be more easily accepted.

6.4 The Analysis Tool

The purpose of the tool is to bring the model to life and create a “one-stop shop” for the scientist to interact and explore data related to CEX. Rather than looking through dozens of documents and spreadsheets, the scientist sees all the available data, including how the model
was trained, and can easily select specific values for parameters and see their impact immediately on the dynamic ranges of all indicators. **Figure 13** displays a detailed example of the interface.

![Figure 12: A snapshot of the analysis tool (values and names masked)](image)

**Table 9. Detailed area by area explanation of the tool snapshot**

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Displays the data about winning model, as shared earlier in this chapter. This information is static and does not change due to user interaction.</td>
</tr>
<tr>
<td>2</td>
<td>Allows selection and filtering of specific values and ranges for different parameters.</td>
</tr>
<tr>
<td>3</td>
<td>Displays applied filters and allows changing range or removing filters.</td>
</tr>
<tr>
<td>4</td>
<td>A Graph between an indicator and a parameter based on the filtered data</td>
</tr>
</tbody>
</table>
Dynamic ranges of indicators, effected by current selection. For example, the dynamic range of indicator 2 is 6-30, but the original range of indicator 2, as shown in area 1 is 5.2-34.

Dynamic ranges of parameters, same as area 5 but for process parameters.

**Figure 14** displays another worksheet where the scientist can examine the data sparsity table to assess the analysis.

![Data Sparsity Table](image)

**Figure 13. This part displays the data sparsity table, as shown in section 3.2.5**

**Evaluating Ranges:**
An important part of Process Characterization is to evaluate ranges of parameters and indicators as the company is preparing the scale up the manufacturing process. The simulation tool presented here allows a scientist to interact and dynamically change parameter ranges and to measure their effect on the predicted range of multiple indicators at the same time. It also allows the scientist to select a desired range for an indicator and observe the acceptable range for parameters effecting the indicator.
7  **Recommendations and Contributions**

This section outlines recommendations for future work and next steps that will continue to leverage the value and success of this work for Amgen. We will go on to describe the major contributions achieved by this work. We will also discuss observations about machine learning initiatives and data governance challenges that should be addressed in order to enable extracting the highest value from similar initiatives.

7.1  **Future work: expand research**

Additional opportunities to further this work are in discussion. The models and tool created throughout this work were built as a proof of concept. While work processes, code efficiency, scalability and user interface can be improved, it would be far more interesting to apply the general framework presented here to enable expanding the work to new areas to create more exhaustive models of biologics manufacturing.

7.1.1  **Expand to other unit operations**

As demonstrated here, this approach could be expanded to other unit operations. Biologics manufacturing include different stages in both upstream and downstream processes. The framework was developed with an agile perspective in mind, expecting new unit operations to follow. It would be interesting to see if other machine learning models will be preferable in different unit operations. Creating a model for additional unit operations will directly benefit the process characterization work of process development scientists working on new projects. The two major challenges for this type of expansion will be:

1. The data gathering aspect which we will refer to later in this section.
2. Researching whether the unit operation history and its future stage make sense for the model to draw predictions based on past history. On the upstream area, single-use technologies are changing the scale up concept. On the downstream area, manufacturing of biologics is going through many changes, refinement and improvements.

Both these challenges require support from an experienced process development scientist.

7.1.2 Model Unit Operations dependency

The manufacturing stages in biologics manufacturing are dependent on one another, where the output of the first stage would be the input to the second stage, and so on. This work dealt with only one Unit Operations - Cation Exchange Chromatography (CEX).

This work should be expanded to create a holistic model of biologics manufacturing, across different stages. Creating such a simulation, also called digital twin, would create benefits for Amgen, by allowing accurate predictions, saving of resources and thus accelerating development time.

Although a full digital twin does not seem immediately feasible due to the challenges specified in the previous recommendation, creating a dependent model to incorporate more than one unit operations stage would be the logical next step in the journey to realizing the potential.

7.1.3 BiTE® molecule evaluations

The model in this work contained data of one type of molecules called monoclonal antibodies (mAbs). As Amgen develops new molecule constructs, such as the fusion protein called Bispecific T cell Engager (BiTE), it would be exciting to see how the model performs on different types of molecules. It would also be interesting to research whether including both types of molecules in a single model would improve its performance.
7.2 PD and AMGEN Recommendations

As seen in the above future work, there are abundant opportunities to further benefit from this work. Here we will outline specific observations and recommendations for Amgen and Process Development (PD).

7.2.1 Ensure data governance and Model Governance

In the last few years, Amgen has been able to centralize and democratize its data and the associated analytics, in order enable a faster deployment of data projects and decision making tools. Data from processes and systems that were siloed and not ready for analytics were made readily available for thousands of scientists. New platforms to enable faster development of machine learning projects are currently being deployed.

Data governance has come a long way since the early days of the digital age and new data sources and dashboards that are introduced are being verified before general usage. The appetite for data-driven decision making tools has created proliferation of machine learning projects. The models developed during those projects bring with them new questions. Those questions are out of scope for the work presented, but answering them is of paramount importance for any company that wishes to harness the power of machine learning successfully and efficiently:

1. How to define, classify and maintain machine learning models? How to ensure reusability of existing models to avoid duplication of efforts in development?

2. Why was a certain machine learning technique or approach chosen for a certain model?

3. In what context is the model being used, what decisions or processes is it supporting?
4: What data is being used in the model and where is it coming from? This question is true for every analysis but the answer for it is getting increasingly complicated.

5: Are the model’s results explainable? Should they be?

6: How are the models performing? Which metrics are used? What is the effect of new data refresh on performance?

8: How is each machine-learning model trained?

9. Do the results of the model need to be reproducible? Can we re-run the model in the same conditions and same data? If not, what are the computational and technical consequences of meeting reproducibility requirement?

10. Who is the owner of a model and who is responsible to maintain it?

Although this list of questions might look intimidating at first, the topics they address should be discussed upfront or they will come up in a different way.

### 7.2.2 Datalake paradox – improving centralized data projects

As stated in previous section, Amgen has centralized and brought together data from a wide range of systems and sources to create a data lake. Data is constantly being refreshed from process development, quality control, and manufacturing systems. New data sources are introduced regularly in an ongoing effort.

Recognizing that the complexity of biologics manufacturing is hard to overstate, that complexity is clearly reflected in the data lake as well. The availability of a variety of data, both structured and unstructured, is mesmerizing.
However, in order to import new data source, the source has to be documented, verified and mapped to existing data structures. The data might need to be filtered and converted. The effort to add and maintain data sources is significant and requires expertise. Therefore, not all data sources exist in the data lake. In addition, data exist in different levels of analytics readiness, one example is the availability of documents vs structured, normalized data.

An interesting observation was to watch two very distinct perspectives; the technology personnel was claiming success by pointing to the availability of X% of records/sources for a certain topic, while the scientists were concerned with the integrity of the data and its totality.

Due to the strict regulations in pharmaceuticals, the high standards of compliance and research, have the potential to create a misalignment on what a “ready to analyze” dataset is. This in turn can lead to non-accurate estimations of effort required to research a certain topic.

Internalizing and discussing the disparity between those perspectives of different stakeholders is essential to optimize and manage machine learning projects in an enterprise.

7.2.3 Develop skills for maintenance and upgrade

As machine learning initiatives become more prevalent, the developing of skills of machine learning understanding for a research workforce is unavoidable. Traditionally, scientists do not have the set of skills to interact, interpret, maintain and later upgrade machine learning models, and as the primary users of those models, that will have to change.

7.3 Major Contributions

7.3.1 Multi-Product / Multi Parameter / Multi Indicator Model
The combination of Process Characterization studies data with molecular descriptor data based only on machine learning yielded a multi-product (molecule), multi-parameter and multi-indicator model. No specific algorithm has proven to be the best for all Indicators, but complex models such as Regression Trees and XGBoost were generally more successful.

Using the metric of RMSE / Range Size, out of 22 Performance Indicators, top 17 have a score lower than or equal to 20%, top 8 have a score lower than or equal to 10%. And the bottom 5 indicators had 71 data points or less.

The model enables for the first time to assess and quantify the impact of parameters on indicators, even if they were not tested originally for a specific molecule.

7.3.2 Developed a user interface

Currently, using Prior Knowledge Assessment is manual, static and time consuming process. Scientists have to navigate between multiple systems and dozens of files. We were able to create a user interface that allows scientists with limited machine learning expertise to interact with the model. Scientists can generate predictions, simulate ranges for multiple parameters and indicators, and assess the performance quality of the model. They can also view historical data about Process Characterization studies.

The approach presented does not replace process characterization; it helps to add precision to an existing business process and provides a much faster “one-stop shop” for the scientists. Adding new data is quick and facilitated through the tool. Due to the low quantities of data, as more data will be added to the model, the potential for performance improvement is substantial.
This page intentionally left blank.
8 References


9 Glossary

BiTe® – bispecific T-cell engager
BD – bioprocess development
CEX – cation exchange chromatography
CHO - Chinese Hamster Ovary
CV – coefficient of variance
DOE – Design of experiment
EDL – enterprise data lake
ELN – electronic lab notebook
KNN – K-nearest neighbors
mAb – monoclonal antibody
MOE - molecular operating environment
MSE – mean squared error
NOR – normal operating range
OP – Operating Parameter / Process Parameter
PC - process characterization
PD – process development
PIR – process impact rating
pI - isoelectric point
PLS – partial least squares
PKA – prior knowledge assessment
RMSE – root mean squared error
SGD - Stochastic Gradient Descent
SVM – support vector machines
UI – user interface