Scientific Publications

Journal of Artificial Intelligence and Big Data, 2019, 1, 1256
www.scipublications.com/journal/index.php/jaibd
DOI: 10.31586/jaibd.2019.1256

Review Article

Predictive Analytics in Biologics: Improving Production Outcomes Using Big Data

Kiran Kumar Maguluri [1*], Venkata Krishna Azith Teja Ganti [2]

[1] Sr. System Architect, USA

[2] Support Engineer, Microsoft Corporation, Charlotte NC, USA

*Correspondence: Kiran Kumar Maguluri (kirankumar.maguluri.hcare@gmail.com)

Abstract: Biopharmaceuticals, or biologics, are a burgeoning sector in the pharmaceutical industry, predicted to reach $239.4 billion by 2025. This unparalleled growth is often attributed to the enhanced specificity offered by large molecules over small molecules. The large size of the constituent proteins necessitates the continuous implementation of big data predictive analytics to elucidate the most effective candidates in the lead optimization process. These same methodologies can be applied, and with the advent of machine learning and automated predictive analytics, this is becoming an increasingly facile task, to the augmentation and optimization of the downstream production processes that comprise the majority of the development cost of any biologic. In this work, big data from cell line generation, product and process design, and large-scale lead validation studies have been used to compare the applicability of simple statistical models against these black-box approaches for the rapid acceleration of enzymes to the pilot plant stage. This research can be expanded upon to exploit the big datasets generated as part of the progression of biologics through the development pipeline to further optimize production outcomes. Over the coming months, data from the project will be used to probe which approaches are amenable to which processes and, as a result, more amenable to various economic simulations. The computed optimization objective for the HIT must include the cost of acquiring, storing, and analyzing data to construct these predictive models, alongside the expected commercial reward of choosing an optimally ranked candidate. In this vein, perspective must be taken in the probable future price, capability outputs, and ownership issues of increasingly sophisticated data analysis software as superstructures become more frequent. It is frequently stated that decisions made to reduce production costs are data-driven, but that is not because more economically or energetically costly experiments or production methods are employed; to truly evaluate production steps, dynamic energy, and economic models need to become more commonplace. Conversion of process quality approaches from large questionnaires, risk analysis, and expert opinion-driven methods to statistical and thus more reliable approaches is an area of future research in analytics used herein.

How to cite this paper:
Maguluri, K. K., & Teja Ganti, V. K. A. (2019). Predictive Analytics in Biologics: Improving Production Outcomes Using Big Data. *Journal of Artificial Intelligence and Big Data, 5*(1), 1256. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1256

Received: September 28, 2019
Revised: November 19, 2019
Accepted: December 21, 2019
Published: December 27, 2019

Keywords: Biopharmaceuticals, Biologics, Big Data, Predictive Analytics, Machine Learning, Lead Optimization, Downstream Processes, Development Pipeline, Statistical Models, Black-Box Approaches, Enzyme Acceleration, Pilot Plant Stage, Production Optimization, Economic Simulations, Data Analysis Software, Dynamic Models, Process Quality, Risk Analysis, Cost Reduction, Predictive Models

1. Introduction

Biological drugs are becoming one of the most impactful sectors in modern-day medicine. Due to their difficulty in optimizing in a clinical setting, traditional chemical drugs are losing favor at a fast rate in comparison. To guide complex biological targets, such as proteins, predictive analytics is taking an ever-increasing role. Predictive modeling using advanced analytics can now take large amounts of data surrounding a

particular sample and make predictions about the production, showcasing the complexity of biological systems. These models are being used thoroughly in front-line pharmaceutical production to drive large business value. The use of big data in predictive modeling techniques is often employed to increase production outcomes and efficiency.

Big data and predictive analytics are used together in the industry to enhance biological production to lower costs and increase efficiency. The continuous demand for biological products, particularly in the pharmaceutical and nutraceutical sectors, drives growth in an exponentially expanding global marketplace. However, the production of high-value bioproducts presents several process challenges and impacts sustainability on an ecosystem level. These challenges can include but are not limited to, low product yield, the high overall cost of goods, poor product activity profiles, and long lead times for process optimization, and are often linked to continuous process drift. While much research on metagenomics has been conducted over the past decade, further work is necessary to put it to practical use. This paper will highlight the requirements to facilitate the realization of a functional Fourth Industrial Revolution infrastructure.

1.1. Background and Significance

Biologics have made significant advancements in recent years. They are defined as biological products manufactured using a living system or living organisms. As such, they have opened the door to new ways to treat and manage diseases. Biologics are targeting diseases on a genetic and molecular basis and have been approved for cancer, diabetes, autoimmune and inflammatory diseases, bleeding disorders, cardiovascular diseases, Alzheimer's disease, and more. The global biologics market has seen rapid growth, and the sales of biopharmaceuticals in the US have reached over $275 billion annually. These new therapies help improve the quality of life and longevity for patients and individuals living with a wide array of previously difficult-to-treat chronic as well as rare diseases. As more biologics come to market, the portion of US pharmaceutical spending on biologics is expected to grow. Pharmaceutical companies looking to enter the market have the challenge of satisfying the increasing demand from patients, healthcare practitioners, and payers looking for new options to treat a wide variety of disease states quickly and economically [1].

Current manufacturing technologies use relatively small bioreactors to produce biologics and are time-consuming, producing low quantities of biological drug substances. Even when scaled up, traditional production technologies have difficulties reaching commercial production requirements in terms of processing time, cost, and quality. The difficulty of manufacturing a biological drug by fermenting cells in a bioreactor limits the ability of pharmaceutical companies to fill this healthcare need. The following study positions predictive analytics as an evolving approach to assist with the transition to integrated biologics production. Through harnessing big data, improved outcomes could shape the future of biologics. The persistence of aging populations and the resulting increase in chronic diseases ultimately support the requirement for large-scale biologics production. A call for innovative manufacturing approaches to satisfy increasing demand and lower production costs reflects the present state of the biologics industry. A recent shift has seen pharmaceutical companies expanding their networks to accommodate the production of the growing biologics market.

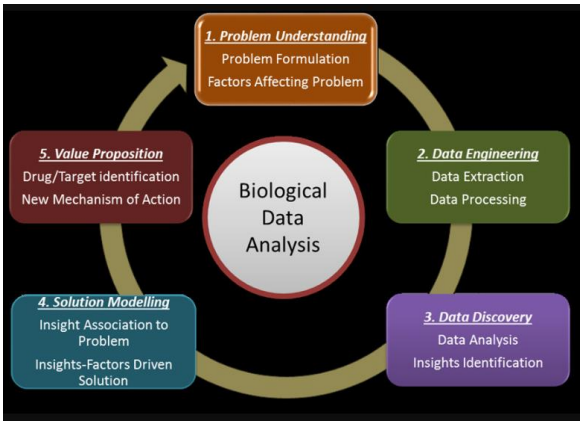

Figure 1. Biological Big Data Analytics.

1.2. Research Objectives

This research study aims to focus on particular research objectives. This will include critically appraising current biologics production methodologies and identifying where the potential exists for improvements in the production outcomes. Once this has been completed, the study will aim to evaluate, using carefully identified case studies, where digitizing the suite of manufacturing steps and using predictive analytics on the big data they will generate is heading and what these leading examples of manufacturing are achieving. These objectives will assist in demonstrating a direct commercial application to underpin the academic findings. The study will attempt to address the following principal research questions through the investigation of specific case studies: What would the value of predictive analytics be if it operates effectively in the future? Are there areas of biomanufacturing that, theoretically, should be doing better than they are? Are there any glimmers of something more promising in our data that need further exploration? This research adds to the body of production scheduling, operational planning, and discrete event simulation knowledge by linking the theoretical work undertaken on operation scheduling and optimization with practical knowledge and examples of the trailblazing industry stakeholders who are adopting and implementing solutions for the digital factory of the future.

As these case studies are based on up-and-coming, cutting-edge predictive analytics in biomanufacturing, their data, conclusions, and recommendations are of value to academia and industry, and this study is the first to map them like this. The digital revolution is upon us as an industry, the economy, and society as a whole. However, there is now widespread consensus that to reach the future digital enterprise using all of these technologies, the key underpinning block is the exploitation of big data and predictive analytic systems. However, this needs to be treated with some caution, as concerns over appropriate and assured data structure along with the other characteristics of big data. Moreover, it is not clear currently what the full and real value of predictive analytics might be. It suggests that the areas of greatest adopter indifference are predictive analytics and recommendations and that considerable value is waiting to be unlocked by companies more fully understanding the predictive analytic systems they are using.

Equation 1: Feature-Outcome Mapping

$$y = f\left(X, \Theta\right)$$

where:

y : predicted production outcome (e.g., yield, purity),

X : feature matrix (e.g., temperature, pH, nutrient levels),

Θ : model parameters.

2. Overview of Biologics Production

Biologics continue to enjoy a surge in popularity due to their potential to provide new therapeutic options for addressing unmet medical needs and serving as critical inputs for new innovative therapies. In contrast to traditional small molecules and other more traditional pharmaceutical products, biologics are large complex molecules of biological origin, typically produced by living cells through recombinant technologies. They generally exhibit higher specificity and effectiveness in interacting with biological targets within a patient's disease pathology. Clinical outcomes and targeted activities may be markedly impacted by even minor variations in the substance properties. It is therefore important that biologics be produced using suitable quality systems and carefully optimized process design. The production of many modern biologics now typically involves a wide variety of upstream and downstream activities. Products first require isolation and capture, such as by mammalian cell fermentation. The captured materials then undergo purification and are fractionated. Polishing procedures may then be applied that could include chromatography or filtering operations. The isolated active pharmaceutical substance may be further characterized and assembled into final dosage forms to be sterilized and packaged. To shed light on methods to balance quality and volume-oriented production issues, levels of detail simulations and integrated modeling tools offer particular promise. As such, predictive analytics may provide particular value for supporting the integration across the production enterprise, particularly the relationships between upstream and downstream operations and quality. This teamwork is particularly essential in chemical process production, often leading to collaboration between functions.

2.1. Definition and Importance of Biologics

2.1.1. What is a biologic?

Biologics are products derived from living organisms. They are complex mixtures of molecules such as proteins, sugars, nucleic acids, or whole cells and can be composed of many parts produced in different biological systems. Biologics are typically much larger and more diverse than small-molecule drugs. Small-molecule drugs, the type most people are familiar with, consist of well-defined chemical molecules manufactured through a series of chemical synthesis reactions. In contrast, biologics are generally produced in living cells such as bacteria, yeast, mammalian cells, or plant cells and not by chemical synthesis. This introduces greater variability in how the product is produced, which can affect the product itself.

2.1.2. Why are biologics important?

Biologics are used to treat chronic and debilitating diseases where treatment by traditional small-molecule drugs is unavailable. This includes conditions like cancer, autoimmune disease, transplant rejection, neurodegenerative disease, infectious diseases, sterility, anemia, HIV/AIDS, hepatitis, and diabetes. Over 400 biological therapies have already been produced and launched in more than 100 countries. Almost a further 900 molecules are currently under development. Biologics were initially considered second-line products due to their complexity and cost. They have since been officially named 'New Treatment' drugs, 'Fast Track' drugs, and 'Orphan Drugs'. Approximately 60% of the 41 NT/FT/OD drugs were biologics. Pharmaceutical companies have at least one biologic in their portfolio, with up to 50% of new drugs estimated to be biologics by 2010. Biologics have extended the treatment options for many diseases and are being used increasingly as the 'first line'. Administered through injection or infusion, biologic products accounted for over $100 billion in global sales in 2015. As disease prevalence increases and patients switch from hard-to-manufacture synthetic drugs, the implication

for manufacturers of the newer protein drugs is that their plants and equipment must produce large amounts of these products to meet increasing demand.

2.2. Challenges in Biologics Production

Biologics are complex therapeutic molecules often produced within living systems, necessitating extensive quality control during production and a high cost for eventual drugs. The small changes in raw materials and the sensitivity to any changes in cell culture can have significant effects on the final product. Moreover, no two batches will ever be identical due to the infinite variability in biological raw materials and the succession of myriad purification techniques needed, and required by law, to detoxify said materials. In these respects, biologics manufacturing is a space where process and product quality are essentially synonymous – the issues with quality control are fundamentally issues with production quality [2]. Regulatory authorities have mandated very perishable timelines for the requirements of a product to be assessed, in addition to intense interdisciplinary testing, which turns many commercial conversations about market flux into the stream of consciousness. The high cell concentrations this process requires result in rapid hydrodynamic changes to the growing colonies which can, in turn, damage the cells and necessitate further repair. This makes extreme scalability – otherwise known as flexibility – a necessity, especially for smaller producers trying to hedge the increasing market volatility of biopharmaceutical commodities. As the production of biologics is a heavily monitored process observed under intense scrutiny on a large industry scale, every incremental improvement is seen as an ethical or economic victory for the wider ecosystem. This has become especially crucial in recent years: one of the most commonly cited reasons for the high costs of even generic biologics comes from the requirement to redesign processes even years after their device approval. The developing answer to these problems is in demand: a faster, more flexible way of ensuring a safe and effective product. 'Decreased time towards market' has become ever more of a noun phrase, and 'less risky clinical supplies' are an added reason for big corporations to fund smaller developers. There is thus still doubtless value in developing faster, less variable methods of control. The most useful of these methods will summarize multiple, replicable process improvements into an inference that can be applied in a standardized way across many different production systems, ideally informing the design of new production setups. In practice, such a standard would require greater flexibility, but we believe it is achievable based on recent biological databases of sufficiently large, high-quality samples.

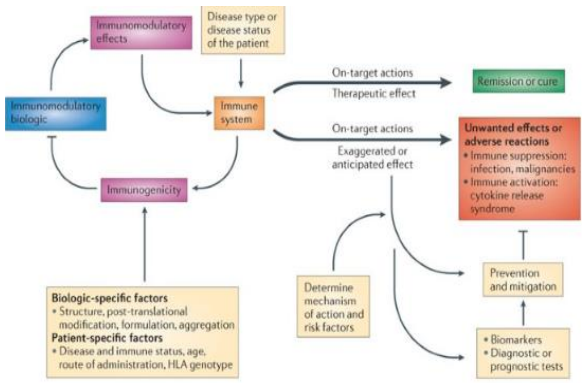

Figure 2. Challenges and approaches for the development in biologics.

3. Predictive Analytics in Biologics

Biologics are a class of medications and treatments that are derived or extracted in some part from living organisms and natural sources. Their production is a complex process that is resource-intensive, capital-intensive, quality-critical, regulated, and subject

to the protection of intellectual property. Ultimately, these drugs are intended to elicit a specific and consistent biological response in the patient, and in many cases are used to treat diseases for which options have historically been limited or nonexistent. In the past two decades, there has been a rapidly expanding interest in predictive analytics in the context of biologics production to enhance predictability and sustain and improve outcomes.

Predictive analytics is the use of statistical algorithms and machine learning techniques to analyze historical data to predict future observations or make informed decisions. This is done in contrast to traditional statistical methods of forecasting, which rely on establishing linear relationships with a large body of data whose principal merit is that of having led to sample averages that can be termed as forecasted results. Predictive models can be developed and employed at multiple levels on a spectrum of detail, from broad-brush mechanistic predictive models operating at the process screening and design space development level to high-resolution statistical models and/or machine learning-based models that may be employed in the operational context.

Using these models and models in between, numerical predictions can be developed for a wide range of process attributes including many other types of information. Although the integration of such predictive information into operational workflows is an active area of research and development, when appropriately applied these predictive insights can improve the overall efficiency of the manufacturing process and enhance the in-line and/or end-product quality control capabilities in several ways. Biology-based predictive modeling can help to modernize many aspects of biological manufacturing.

3.1. Definition and Conceptual Framework

Definition and Conceptual Framework Predictive analytics is the process of recognizing patterns in quantitative datasets and constructing advanced models based on these patterns to forecast future single or multiple variables. The first two major blocks include some requisites for using predictive analytics for real systems and product development. There are three main evolutions of predictive analytics: implementation, spending time, and applying the concept of iteration cycles. In the title of each block, blank lines also explain the need for using the predictive analytical process on nested systems. Data collection is required because predictive analytics relies on the model creation of the current behavior of a system based on past aggregated data. These data must be processed to ensure that the next step is carried out. Data analysis and prediction ratings are processes involved in prediction. The targets are determined based on our aim, objectivity, and study variables. Quality and accuracy are very important in predicting the result. After the rating process, statistical methods or algorithms produce useful insights that enable us to quickly make accurate predictions. The results or interpretations and inferences continuously repeat the process and improve the model as needed until they align with the previous system and form a similar consistent outcome or other evolved output. The conceptual framework of predictive analytics system application towards systems described above also represents a possible systematic or structured methodology for applying predictive analytics to manufacturing operations and systems of biologics. In this period of rapid development in the biopharmaceutical sector, any robust application or proposed changes are incorporated. Consequently, bio manufacturers can benefit from adopting new techniques to bridge a strategic gap and improve the quality and consistency of biologics [3].

Equation 2: Error Minimization in Predictive Models

$$\min \mathcal{L}\left(y, \hat{y}\right) + \lambda \left\|\Theta\right\|^2$$

where

$\mathcal{L}(y, \hat{y})$: loss function comparing true outcomes y with predictions $\hat{y}$,

λ: regularization parameter,

$\|\Theta\|^2$: model parameter regularization term.

3.2. Applications in Biologics Production

Predictive analytics can be used across the entire primary and/or secondary biologics production processes, e.g., from optimizing upstream cell culture conditions to reducing the high batch-to-batch variability in the final drug product due to differences in process extensions to sterilized filtration solutions. At an early development stage application example, predictive modeling was applied to a couple of CHO cell culture processes, depth filters, chromatography, and final UF/DF steps, focusing on the prediction of the final product infectivity in chromatography eluates. These basic models can be extrapolated to potentially develop platform models for the prediction of any biologics product end quality during normal production and with allowed endpoints such as above. By using predictive models for the above-described examples, two powerful strategies can be highlighted: the future market opportunities of real-time release of products, because of the use of more precise quality prediction tools that can utilize real-time process data with more informative value; and how extreme risk reduction that has higher value economic reimbursement may be provided by the production of personalized medicine, especially in oncology. Proactive steps can be taken just in time, like adjusting process flow or adding a filter or other to the process, in such cases to reduce the larger cost impact of a product recall because the system is in control with management decision analysis procedures. By using outcomes from sophisticated predictive and multidisciplinary analytics, this approach can allow treatment sequencing by better predicting on-treatment reactions and long-term outcomes. It allows seamless patient identification on the fly and can optimize trial designs to reduce enrollment of patients for injected treatment. Optimized production strategies, both GMP and QbD, are desired and are under strict specifications to provide the maximum allowed batch-to-batch performance deviation over 3σ. These limited trend lines effectively define multi-batch acceptance criteria for either passive or active comparability studies whenever there is a planned process improvement. In certain cases where the improvement is considered revolutionary, batch one may also be used to compare with the reduced process range specifications because here one effectively initiates a new reference design. More optimized process design necessarily and significantly reduces costs. Potential quick returns on investment, and businesses strategically, include the ability to: (i) deliver superior biologic production over the competition; and (ii) solidify market share.

4. Big Data in Biologics

The increased digitization of biologics production operations is generating large quantities of information. When applied to biologics production processes, the data generated can be defined using the characteristics of big data: volume, velocity, and variety (both structured and unstructured). This data is generated at different stages along the biomanufacturing value chain and through different means, such as batch records, sensor readings, and other quality data. Predictive data is generated as a result, capturing hints of the future evolution of a production process. However, this 'data sea' and the large number of people, technologies, instruments, and data systems that this data touches on its journey to completion present both challenges and opportunities.

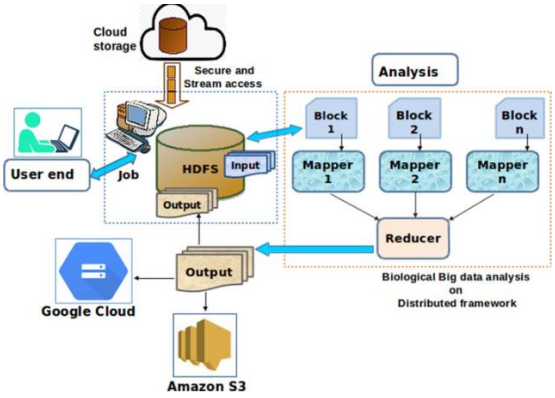

Figure 3. Big data in biology.

Currently, there is no prescribed single method or procedure by which to manage, process, or store this data. Regulatory authorities are prompting control discussions. Utilization of such diverse data sources as well as predictive technologies that bolster their utility is the essence of predictive biomanufacturing. The collection and management of data essential to this end, to ensure wellness in the patient, form the foundation for preventative therapies. In large part, this preventive data arises through high-volume diagnostic testing; when utilized correctly, such data could enable predictions about adverse outcome pathways and modes of action. This preventative data is the core asset of platforms such as the Quantified Self. While different from a production context, this principle applies to biopharmaceuticals and has been described as Quality by Control. The goal of Quality by Design in this context is to make the most informed decision at each point in time given all available information – this is the domain in which we need the next tier of insight, predictive Quality by Design, to successfully navigate the tension between quality and cost/economics. Regulatory controls governing clinical data are predicated on the use of predictive analytics to derive clinical insight.

4.1. Definition and Characteristics of Big Data

The term "big data" refers to voluminous and heterogeneous datasets. Its three defining characteristics include its size, whimsical generation frequency, and unstructured, semi-structured, or structured morphologies. Their sizes range from some terabytes to several petabytes per individual data instance, rapidly mounting with the added number of data dimensions. This exponential sprawl is fueled by various scientific tools that continuously generate new data. When unstructured or semi-structured data are involved, these datasets are challenging to manage and process through regular database management tools. Big data often emanates from surge volumes of digital data, such as instrumentation sensors, weblogs, online transactions, audio and video recording, e-commerce, emails, scientific experiments, surveillance, automated data gathering, social media, internet search, and so on. Petabytes of data per day or even hour, and even single data instances larger than 50 TB are not rare. Big data offers an ungrouped goldmine of information that can be leveraged to unveil new insights and revolutionize decision-making in various arenas, including biomanufacturing.

Presently, commercial biologics manufacturing and most of their supporting services are deluged with big or mega big data inflows due to automation, paperless manufacturing, and direct digital data flow from supporting services and raw materials suppliers. It is believed that the interconnection between big data and biologics production outcomes could untap unutilized opportunities and introduce unprecedented innovation, capability improvement, and efficiency. Apart from revolutionizing biologics, big data offers unparalleled opportunities for streamlining and redrawing decision support systems, with potential breakthroughs in commercial yields, cost reduction

stimulus, regulatory enforcement, adaptive trials, and demands. The volume and velocity of big data protuberate our storage, computational machinery, and security abilities, thus attracting growing interest in dedicated data analytics and novel data-driven frameworks. To circulate big data on the spot and gain instant insights, big data cannot rely on robust data models that are prebuilt. They require flexible exploration and extensively parallel interrogation and mining tools to produce actual data usefulness that is considerably scalable. Furthermore, because big data are stratified in the data chain, all data-oriented knowledge management systems and processes need to be tailored for ingesting, storing, and processing big datasets, alongside their sanctifying needs such as security and audibility, handling rights, and privacy-intensive information in compliance with such data use agreements [4].

4.2. Role of Big Data in Biologics Production

The growing interest in the practical application of big data has seen the generation of vast amounts of biological data, which is being used to inform decision-making and strategies to improve efficiency in the production of biologics. In the production of complex biological products, factors such as alcohol, metabolic intermediaries, temperature, nutrient exhaustion, and the induction or inhibition of regulatory elements contribute to differences across cellular populations. In small molecule manufacturing, these discontinuities lead to multiple impurity populations that can have surprising toxicological responses. More specifically, genomics, proteomics, and other 'omics have allowed for a greater understanding of cellular responses to varying gases, media formulation, and other production parameters, resulting in the optimization of single-use bio-production systems with fewer steps, lower costs, and increased speed to market.

Variability in biological production is significant, and it is more complex with larger datasets. One solution to manage variability involves choosing methods, parameters, and systems that minimize it, for instance, through platform process development in seeking rapid, adaptable, efficient bio-production systems. Operating parameters for cell size, time to begin or end induction, flow rate changes, temperature changes, or other apoptotic-inducing factors can be informed by a significant dataset that has informative parallels. Sensors might detect a change in metabolic flux that would move a process away from a predictive outcome and thus necessitate a change in the overall process of cell inhibition or killing. Concerns about confounding in small datasets over enzyme production and the use of inoperancy in multi-dimensional systems have been raised. Finally, the potential for a large collection of varying data to disrupt or distract production teams from the decision-making process and the challenges faced in protecting the asset that is the data have been discussed. But we will adapt. We always do. Access to big data is changing the production of biologicals.

5. Case Studies and Examples

A range of examples will be discussed to provide insights into how predictive modeling can be used within the biopharmaceutical industry. Four case studies will be used to illustrate the implementation of predictive modeling within the industry and the benefits they have delivered. Three case studies provide details on improvement within the fermentation stage of production: a detailed mathematical model has been put in place to automatically predict the main output categories as a function of process inputs at scale, the at-line data (predicting around 450 analytes), and results from final physicochemical assays (such as the amino acid profiles and UV spectral information). Although a detailed AI model did not replace the knowledge of primary experts, it did take care of a lot of detail, guiding the main steps. It also helped to identify which fermentation stages needed to be tested at scale, followed by wet chemistry methods. The high-quality process (in terms of protein quality and quantity) now in place benefits from the implementation of predictive modeling within the biologics area.

The implementation of predictive modeling is broken down into the following stages: development of new process development cleanse predictions for QA release stages, predictive glycans at-line versus QA release, predictive protein quality analytics at-line versus QA release and drug release, scale-down surrogate development, lab scale transfer, and focus on what are likely to be other for detailed composition work including release, at-line, and scale expansion to assay update. To maintain predictive robustness, the control of the feed variables is limited initially. The data captured from this study is also representative of the range of processes being developed. The modeling study indicates where the model developed the process. The process with glutamine was developed first, followed by the new feed medium predictive gradient cycle, showing an increase in development cycle time. Within further development runs, the model ensures the quality of the elapsed to predict the drug release assays remains the same, even after seasoned feed stage time was put back in. The model has predicted the stage where it will switch. Now a significant percentage of values predict that it will switch at about the same stage to levels. The model is not yet predicting the correct media feed stage time for the shift-down factor. Overall, the models currently under development predicted the correct finishing point, including the feed stage, in a majority of the runs. The main change that has been made to the cascade API is the movement to a specific point at the end of every growth phase.

5.1. Successful Implementations of Predictive Analytics in Biologics

In this section, a few successful stories of applying predictive analytics in the context of biologics production are considered. These stories reflect the diverse contexts in which predictive analytics can be applied and the variety of sophisticated methods that can be brought to bear on practical biologics production problems. They serve as both examples of how to effectively use these methods and as a rationale for their broader adoption.

In the case of a biotherapeutics manufacturer, a predictive model was developed for use in routine monitoring at lower management control. Predictive modeling of product quality was performed on rich data sets obtained during a fed-batch bio-processing of an antibody fragment. The multivariate calibration model generated with PCA-Linreg could predict 20 soft sensors correlated to physical and chemical product properties using penetration of online production data. The offline final soft sensor model could predict the time profiles of these attributes in new data sets. All applications show that this modeling approach was able to predict the relevant quality attributes with an accuracy sufficient for real-time process steering or as a fault detection tool [5].

The use of time series modeling to improve the efficiency and cost avoidance in batch biopharmaceutical production demands is showcased. A commercially important flocking step from fed-batch Chinese Hamster Ovary cell culture was studied to determine the profile of precipitation kinetics and offline CF values. 178 fed-batch data points were modeled with an approach that couples the COBRA framework with neural network models to predict the flocking kinetics. The resulting model was validated with several production runs. When using the predicted flock size, significant improvements in purification steps were shown. Estimates of both treated chromatographic column surface area and purified production are presented as this lower management control approach. Efficient forecasters that depend upon spectral data were developed as well. These were then used to make data online with an acceptable accuracy. The development of real-time flock forecasting in online mode is also discussed. These findings show that the modeling approach has positive implications for the ability of an enterprise to perform strategic asset management using real-time process knowledge and thereby can lead to cost avoidance. The high performance of the model is then demonstrated for a set of one hundred fed-batch cells prepared and processed over a few weeks to validate it in a real-time mode further.

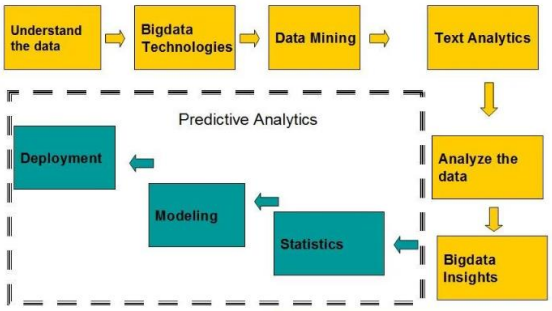

Figure 4. Big data Analytics and Predictive Analytics.

6. Challenges and Future Directions

Notable interest in predictive analytics has emerged in the biologics production industry as facility and process complexities have increased and the demand for biologics has grown. We reviewed what predictive analytics are required for biologics production and where the practice of predictive analytics is relative to these requirements. The steps in predictive analytics, the relevant software and database tools, specific examples, required infrastructure, and limitations were explored.

As interest and capabilities for the practice of predictive analytics increase, several challenges remain, including data integration complexities, workforce capability hurdles, and regulatory hurdles. The industry is only beginning to understand the potential impact of predictive analytics on biologics production outcomes. Most predictions in use are relatively underpowered and see limited application; however, adoption serves as a stepping stone into advanced predictive capabilities. Methods that may be required in the future will include advanced multivariable analysis, better methods of data collection and unification, and a better understanding of the statistical modeling process.

Some of the most significant challenges facing the widespread use of predictive analytics include regulatory requirements, the need for standardization of processes, enhanced technologies and methodologies that can predict a wider variety of outcomes better, and better data management systems. One limitation of current capabilities is that predictive modeling related to product yield is typically correlated with proximal process measurements and therefore directly influenced by the process, which may not provide the best opportunities for learning more about the system. In the manufacturing industry, advanced predictive models contribute to system optimization and are valuable, even by the virtue of enabling the analysis of risk data. Thus, the industry should not dismiss early, currently underpowered applications.

6.1. Current Challenges and Limitations

While the desire is to embrace advanced analytics such as predictive analytics to extract value from manufacturing data and achieve optimal production, established forms of advanced analytical tools rarely go beyond data visualization and reporting. This is not due to a lack of interest; rather, it is due to several challenges. For instance, processes in biologics production occur across different labs and instruments, but data standardization and integration across platforms are uncommon, resulting in suites of tools and technologies with autonomous data silos. Additional stumbling blocks include new skill needs in the workforce for data analytics and predictive modeling, as well as data quality and security concerns.

With such limitations in mind, the industry continues to be cautious about the level of data that should be shared and analyzed. Regulatory environmental challenges concerning the validation and long-term maintenance of pioneering analytics approaches are also significant. Because of these aspects, the business process and organizational elements of implementing predictive analytics are the most often cited reasons for the

limited adoption of predictive analytics across the biomanufacturing segment. Finally, statistical modeling has not been embraced by the bioprocessing field as quickly as in other industries, partially owing to the complex behavior of large, suspended cells, media interactions, and expensive, high-stakes bioreactor trial and error.

Equation 3: Big Data-Driven Optimization

$$\hat{X} = \arg\max_{X} \left[Q_{quality}(X) + \alpha Q_{cost}(X) \right]$$

where

$\hat{X}$: optimized process parameters,

$Q_{quality}(X)$: quality score for biologic production,

$Q_{cost}(X)$: cost-efficiency score,

α : weight for cost-effectiveness.

6.2. Future Trends and Innovations

Facilitated by advancements in artificial intelligence, predictive analytics methodologies are going to become more intelligent. It is expected that machine learning and deep learning algorithms will facilitate advanced predictive modeling. Big data-focused predictive analytics combined with traditional production knowledge and experience will provide an opportunity to develop robust predictive analysis capabilities for production IQAs. Generating predictive analytics using real-time data will support in-the-moment data analytics, enabling operational decision support. This is predicted to be supported by advancements in big computing and real-time data analysis tools. Furthermore, predictive analytics will focus on not just predicting failures, but suggesting operational changes to prevent the failures from occurring. This will be an enhancement to the current capability most predictive analytics systems have for providing preventative analytics.

Future advances for predictive analytics platforms and methodologies may also include the implementation or consideration of technologies such as OEE, simulation models, or digital twins, for optimizing the knowledge gained from predictive analytics tests or for developing performance indices for multivariate and multi-stage datasets. With the increased use of digital data in bioprocessing, it is expected that a range of different industries will provide advances in predictive analytics methodologies, such as the finance or semiconductor industry. This will provide options for cross-industry working and the development of novel approaches to predictive analytics for production IQAs. Future advances using big data in this area will be included in a new report on Big Data in the Manufacture of Biologics, to be released within the next year. Finally, the increasing body of research focused on this area may ultimately lead to the development of predictive analytics performance frameworks for in-process IQAs in biomanufacturing. The development of such frameworks, however, may be subject to rigorous scientific scrutiny and is a longer-term milestone for this area of research.

Figure 5. Challenges and Limitations of Predictive Analytics.

7. Conclusion

In sum, the purpose of this essay has been to make the case for the deployment of powerful big data-driven decision-making solutions, including predictive analytics, as a means of deeply enhancing production outcomes in biologics. The evidence in support of this belief is robust – big data in particular is a fast-growing area of investment, and predictive analytics is already contributing to significant improvements in bioprocessing outcomes – including raising titers, yield, and paving the way for the onset of next-generation biologics. Throughout the text, the significance of these findings was expanded upon, presented in terms of an urgent call to apply proactive, expert-utilizing solutions in biologics production to finally meet the world's huge unmet protein needs. Big data and predictive analytics are capable of enhancing the efficiency of biomanufacturing in complex ways. The initial costs of predictive analysis programs tend to be high, but the true costs can eventually be quite low if the cost of failed biologics discoveries is taken into account. In the future, other precious bioprocessing parameters such as titer levels, yield levels, and the freshness of vital cells could be predicted by machines better than by human experience, and a company is already pioneering the use of predictive analysis to sharpen predictions and develop new high-value biologics. Effectively operating biologics discovery and production lines is fundamental to creating value in the future of medicine, biotech, and pharma and ensuring that great biologics are not, lamentably, too rare. This text should be read as a clarion call to stakeholders at every stage of drug discovery and design to study the possibilities of deploying predictive analytics to engender better outcomes.

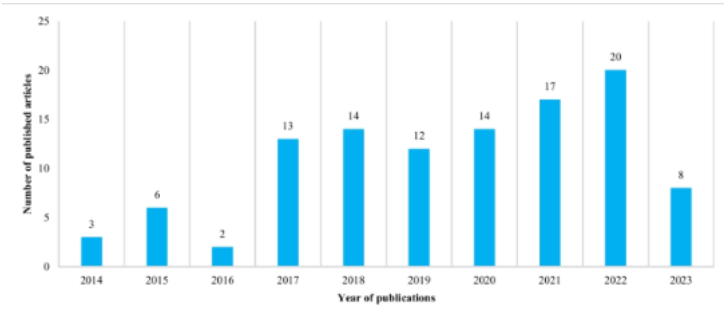

Figure 6. Graph of Big data and predictive analytics.

7.1. Summary of Key Findings

Predictive analytics in the production of breakthrough therapies can significantly increase the efficiency and quality of the final therapies produced, fast-track production launches, and mitigate supply chain disruptions leading to high product loss post-approval. Both cost savings and high-quality real-time and post-process big data are needed to fuel impactful predictive computational models and analytics. The generation of such data needs to be complemented with the coordinated adoption of data analytics technology and predictive computational capabilities. Technical and organizational silos both need to be broken to allow for real-time decision-making and the steps needed to implement new methods and processes, ultimately incrementally transforming manufacturing facilities. Lessons learned from those implementing predictive analytics now indicate that they are ready for change and able to meet the challenges of model validation, data cleansing, and aggregation. They are also finding the same predictive models useful for strategic organizational decision-making in upstream process mode design and have newly configured models to help proactively address future challenges in predictive computational facility and equipment monitoring.

An increasing churn in new biologics entering pipelines, coupled with the steady rise in the cost to manufacture existing products, is necessitating new strategies and

operations for the manufacturing of biologic therapies. A modern facet of production forming the basis of a future biologics facility is the concept of predictive manufacturing: integrating big data into a production environment to drive efficient and timely production outcomes aligned with therapeutic quality attributes and specifications. Prediction and subsequent mitigation of production uncertainties and deviations are especially paramount in the multivariate, complex world of biologics manufacturing and human processes.

7.2. Implications for Industry and Research

Implications for Industry and Research: The findings in this study suggest that competitive biotech and pharmaceutical companies will start offering such services shortly, exposing yet another biologics sector in this market to the latest in advanced analytics. Stakeholders who do not have predictive analytical tools at hand may soon become laggards from a business point of view. Additionally, the outcomes and insights presented offer a new direction for research, producing the data required to design, verify, and validate predictive models. Much remains to be developed and verified on this predictive model, as it was based on many assumptions enumerated in the earlier section. Further collaboration with industry partners who have access to real-world plant performance and production data will provide the necessary empirical evidence. Opportunities for further research also include the development of a predictive framework that considers other data elements such as input costs, raw materials, and utilities that were outside the scope. Additional predictive analytics for different OSI system layers will be introduced shortly.

In the domain of data, academia appears to be centered at the core of this framework. Indeed, key scientific challenges such as exploring new production materials and products and new data-based automation and cost reduction are to be gained from new knowledge and offered as an opportunity for profit to the so-called industry 2.0 (biotech industry). Therefore, close collaboration between academia and industry offers great potential for increased added value and the rapid transfer of knowledge from industry to academia. It should be noted that the use of big data and cyber-physical systems has ethical implications such as data privacy and the ownership and responsibility for the analysis results. In general, the need for predictive data in the biotech sector opens up new research directions. It has recently sprung up with the promise to lead to a better understanding of the complex dynamics that occur in bioreactors through statistical learning models and the collection of large data volumes linked, for example, to daily recorded physiological time series measurements. However, there is still significant demand for data from the biotech industry of the future: data from processing plants for bio-industries.

References

[1] Vankayalapati, R. K., & Rao Nampalli, R. C. (2019). Explainable Analytics in Multi-Cloud Environments: A Framework for Transparent Decision-Making. Journal of Artificial Intelligence and Big Data, 1(1), 1228. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1228

[2] Dilip Kumar Vaka. (2019). Cloud-Driven Excellence: A Comprehensive Evaluation of SAP S/4HANA ERP. Journal of Scientific and Engineering Research. https://doi.org/10.5281/ZENODO.11219959

[3] Chintale, P., Korada, L., Ranjan, P., & Malviya, R. K. (2019). Adopting Infrastructure as Code (IaC) for Efficient Financial Cloud Management. ISSN: 2096-3246, 51(04).

[4] Syed, S. (2019). Roadmap For Enterprise Information Management: Strategies And Approaches In 2019. International Journal Of Engineering And Computer Science, 8(12), 24907-24917.

[5] Mandala, V. (2019). Optimizing Fleet Performance: A Deep Learning Approach on AWS IoT and Kafka Streams for Predictive Maintenance of Heavy - Duty Engines. International Journal of Science and Research (IJSR), 8(10), 1860–1864. https://doi.org/10.21275/es24516094655