APPLIED MATERIALS® | Applied SmartFactory®

Technical white paper

Science Guided Machine Learning (SGML)

Integrating Mechanistic, Empirical and Hybrid Models Through “Expert Unfolding”

Contents

Science Guided Machine Learning (SGML) Integrating Mechanistic, Empirical and Hybrid Models Through "Expert Unfolding"

Lucas Vann, Roy Marsten, Yoram Barak

Abstract

In the world of biopharma manufacturing, batchwise unfolding techniques are commonly used to develop batch process models. However, these techniques result in converting the time series data structure from multiple rows of data with few columns (each column representing one process variable) to one row of data per batch with thousands of columns (each column a time point value of each variable). In order to model this data structure, Partial Least Squares (PLS) or other common techniques must be used to reduce dimensionality of the data set. However, one main drawback of PLS is that it is a purely data driven model and therefore requires a high number of batches to yield accurate predictions. PLS does not leverage mechanistic understanding which, often, has been developed over many decades by engineers and domain experts. Taking advantage of data integration and advanced modeling with AI/ML, we present the development of a different unfolding method in biopharma processes, using Applied Materials SmartFactory Rx® solution platform. This 'Expert Unfolding' framework employs process understanding, which can unlock the use of mechanistic and hybrid modeling in a Science Guided Machine Learning (SGML) approach enabling optimization methods for enhancing process development, quality control, and increasing yield.

Background:

The main goal of any manufacturing industry is to produce a product within prescribed quality specifications. The ease with which this objective is met is directly related to the complexity of the product in conjunction with the ability to adequately control the way in which it is manufactured. Biopharmaceutical production, unlike traditional medicinal products manufactured using consistent chemical and physical techniques, involves biological processes with nonlinear dynamics, inherent batch variability and high sensitivity to minute changes in environmental parameters (Ündey et al., 2010). In addition, raw materials that can be extremely complex are often variable in composition, which can have an unpredictable and substantial impact on cellular metabolism (Read et al., 2010). Cellular growth and product formation in a bioreactor is recognized as the most complex and significant unit operation in manufacturing a biopharmaceutical and governs the success of the overall process. However, very few sophisticated analytical measurements are performed *in situ* and only a handful of critical parameters such as pH, dissolved oxygen (DO) and temperature are commonly monitored in real-time (Chopda et al., 2016).

The clear need to increase process understanding and control resulted in the Quality by Design (QbD) initiative. Soon after, the FDA realized that the advanced control required to ensure quality would not be possible without adequate and reliable monitoring and, as such, the Process Analytical Technology (PAT) initiative was born in 2004 (Izat et al., 2014). The relationship between product quality, cell metabolism and environmental Critical Process Parameters (CPPs) can be monitored closely through determination of in process Key Performance Attributes (KPAs). It is well understood that the ability to monitor CPPs is paramount in developing the required process understanding that enables the advanced process control necessary to achieve enhanced quality in a consistent manner (Biechele et al., 2015). However, with the variability of incoming raw materials as well as slight changes in seed expansion and growth conditions in addition to the limited monitoring available, there is a dire need to develop actionable understanding through data analytics and modeling.

With the increase in PAT and advanced monitoring taking place in the industry there is now a growing challenge of how to make best use of the data that is generated and transform it into process knowledge and understanding (Marx, 2013). It has been noted that a large part of the future of quality improvement in biomanufacturing will be accomplished by better data analytics of the monitoring that is already in place making possible more advanced control (Langer, 2013). The goal is, therefore, to identify meaningful data that will lead to process understanding, which ultimately enables process control. This advanced control is based on the link between process knowledge and product quality that is provided through advanced data analytics and ensures a more robust overall process (Rios, 2014). In order to perform data analytics, the challenge of data integration from multiple sources must first be overcome. In many cases data is generated and stored in different locations based on the technology being used. Standard bioreactor data is often stored in a Supervisory Control and Data Acquisition (SCADA) or a Distributed Control System (DCS) while inline monitoring using NIR spectroscopy or online off gas analysis would be stored in another location and can often even be of different data types based on manufacturer software. An integration tool is paramount to enabling the analysis of all types of data simultaneously to build the optimal multivariate data analysis (MVA) models for enhanced process understanding and statistical process control (SPC). Data integration between multiple sensors from different manufacturers is still a large challenge today and is a requirement for advanced process control (Graham, 2016).

The use of multivariate models to generate “soft sensors” where quality is inferred from process measurements has been in effect for a number of years, however, there is a major challenge in the ability to execute those models with live data and implement process change in a real-time manner, specifically in a manufacturing setting (Hausmann et al., 2017; Mandenius and Gustavsson, 2015). In addition, most biologics modeling is empirical, or data driven, which results in a poor ability to extrapolate beyond the data set utilized to build the models. Current focus has been on developing mechanistic models around typical mass balances for viable cells (Kyriakopoulos et al.,2018 & Yahia et al., 2021) or energy balance models using metabolic pathways (Quiroga-Campanoa et al., 2018). These models are much better suited for extrapolation if the process were to shift outside of previously known conditions. However, cellular metabolism is extremely complex

and variation is often too difficult to predict using mechanistic models alone. Therefore, the best model approach is to use hybrid mechanistic-empirical models to most accurately represent the process being studied (Mayalu and Asada, 2014). Combining science-based models with data-based models in a guided approach is referred to as hybrid Science Guided Machine Learning which has great potential for bioprocess improvement and optimization (Sharma and Liu, 2022)

Advanced control strategies require a platform that can integrate data from standard process parameters and any number of external analytical tools in order to execute models generated with applications such as Matlab, R or python in real-time along with MVA and predictive models (such as those developed using PharmaMV) and then utilize the results through control logic that is able to feedback into the process. In addition, the platform would need to have the capability of managing alerts and alarms with contextual information and ensure that the correct knowledge was delivered to the right individual at the ideal time. This advanced process control would need to be integrated in such a way as to be able to adjust set-points in existing Proportional, Integral, Derivative (PID) controllers that may be under local Programmable Logic (PLC) or Distributed Control (DCS) (Rios, 2014). Full integration of data sources and data systems utilizing a platform that can enable virtually any type of modeling application to execute in real-time does not currently exist in a manufacturing setting.

Platform Description

Our Applied Materials SmartFactory Rx suite of solutions brings analytics, maintenance and scheduling/dispatching together to optimize process and resource utilization. Deployed end to end across the value chain from manufacturing level to enterprise level transforms your plant to an agile, data-driven environment that supports intelligent decisions from shop floor to top floor, as shown in Figure 1. Our platform provides world-class features and algorithms for advanced real-time process analytics including Big Data, mechanistic modeling, closed-loop and model-based control, real-time adaptive scheduling and rule-based dispatching, and machine learning for predictive & prescriptive maintenance. The SmartFactory Rx platform aligns with a number of pharma initiatives, such as 6 Sigma, Lean Manufacturing, QbD, PAT and Continued/Ongoing Process Verification.

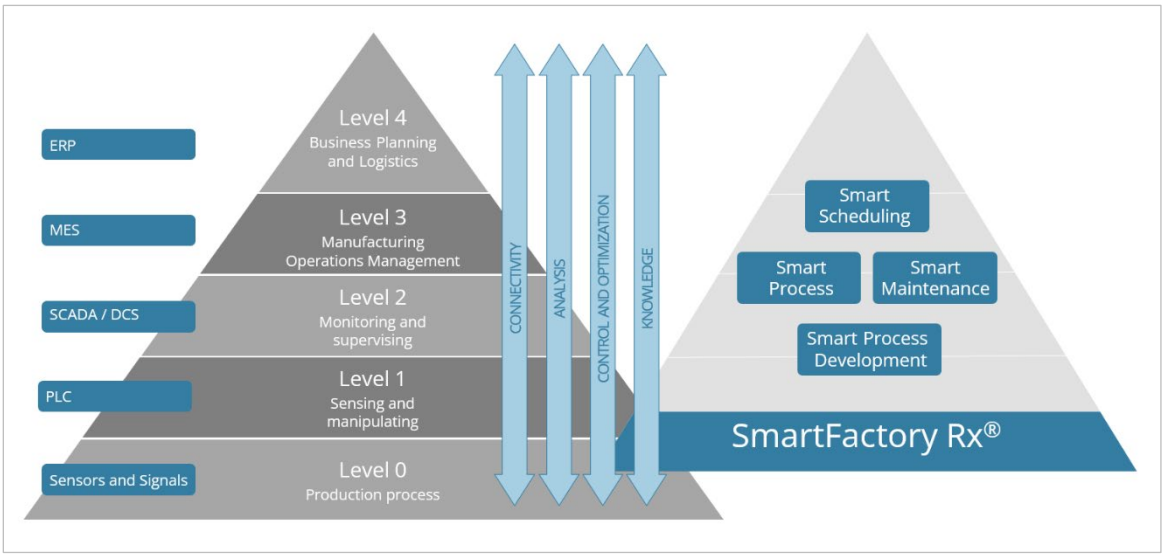

Figure 1 - SmartFactory Rx Platform by Applied Materials

SmartFactory Rx consists of four domains: Smart Process, Smart Process Development, Smart Maintenance, and Smart Scheduling all which inter-operate with other systems/applications from the manufacturing to enterprise level to enable Knowledge Management. Smart Process and Smart Process Development have been designed to collect data from disparate sources, aggregate, contextualize and analyze the data, across the value chain, from raw materials to multiple unit operations and to finished products, enabling advanced process monitoring and control strategy development for process and equipment health optimization. Smart Process interoperates with existing sensors, equipment, PAT analyzers, lab instrumentation, process control systems, Manufacturing Execution System (MES), Computerized Maintenance Management System (CMMS) and enterprise business applications supporting data-driven decisions.

Unique Modeling Methodology

Our approach is an equation-based system modeling method executed in real-time or at run-time using a unique strategy engine enabling integration of various data silos to yield optimal results. Our modeling solution technique first employs “Expert Unfolding” where specific phases of the batch are identified and windowed based on features that will be calculated within and across these windows as identified by process and equipment subject matter experts. This purpose-driven segmentation of the batch, seen in Figure 2, then allows for the creation of model equations that can be based on first principles or mechanistic interactions between parameters. These are often related to both equipment and process performance as well as empirical input/output machine learning correlations. Data can be transformed using complex mechanistic algorithms for modeling continuous data streams, sliced in relevant time intervals (usually no more than 10 for a given batch) to determine more complex statistical performance and/or analyzed using a combination of the two.

Figure 2. Expert Unfolding Windows in Real-time Dashboard with "Golden Batch" limits

In contrast, ordinary batch-wise unfolding uses a regularly spaced grid of small time intervals. In ordinary unfolding, each independent variable (raw or soft sensor) is replaced with a collection of variables, one for each time interval. This results in a 2-dimensional matrix with a row for each batch and a column for each (sensor, time interval). If the process is long, say several days, and the time steps are small, say one minute, then the number of columns can be very large. Expert unfolding, on the other hand, creates a variable for each (sensor, window) combination. Because there is a small number of windows, there is not an explosion in the total number of variables. In practice, the partitioning into windows can be done on the values of some process variable rather than time.

The second major advantage of Expert Unfolding is in the creation of soft sensors. The windows determined by the expert, because they are the natural stages of the overall process, are a more natural environment for defining soft sensors that are relevant in a particular window. These soft sensors can be based on the underlying physics, biology and chemistry of what is happening at that stage. With a uniform time grid, soft sensors have to be defined over the whole duration of the process.

The third benefit of Expert Unfolding is the use of summary statistics. Ordinary unfolding uses the value of each sensor at the end of each time interval. This final value is just one of many possible summary statistics that can be computed over a time interval. See Figure 3 for an example of possible summary statistics. Again, the process windows are a more natural setting for an expert to decide which summary statistics are useful for each sensor in each window. Very specific statistics are employed in order to describe the raw data and soft sensor trajectories within the window as well as identify if the raw data or soft sensor traces are deviating from normal operation which would result in a loss of quality or reduction in yield. This Expert Unfolding and feature generation by application of the aforementioned statistics such as kurtosis, hit counts and quartiles etc. need

only be configured within the software and no coding is required. Limits for these data transformations, univariate statistics, first principle as well as mechanistic models, are determined based on statistical variation of good performance. Any deviation in performance can then be traced back to previously identified failure modes.

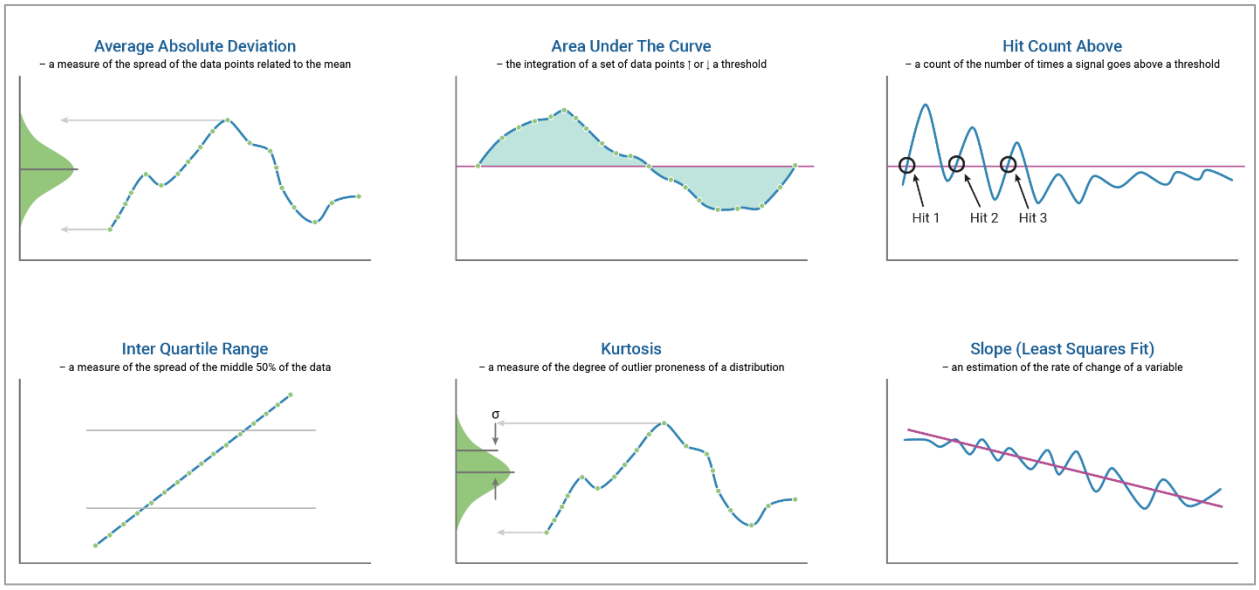

Figure 3. Subset of native statistics that can be selected for feature generation to unfold batches

Machine Learning algorithms are then applied to the generated Expert Unfolded batch data in order to predict and optimize a variety of unit operations providing prescriptive actions to operators at opportune times within validated limits. This SGML methodology has been demonstrated in a wide variety of large and small molecule pharmaceutical manufacturing processes specifically in bioprocessing examples such as fermenters and bioreactors where the models have already shown improvement in yield and quality for many top pharma companies. These hybrid SGML models determine if culture performance attributes are varying and predict whether a process change is required to maintain optimal growth and product formation while limiting negative by-products all within the context of the equipment operation performance. Figures 4 and 5 are examples of how SmartFactory Rx provides analytics visualization as part of the engineering interface and a web-based dashboard. This user-based visualization enables personnel at every level across the organization to understand the health of the process (Figure 4) and the prescriptive action that needs to be taken for a specific unit operation (Figure 5) to optimize growth and product formation.

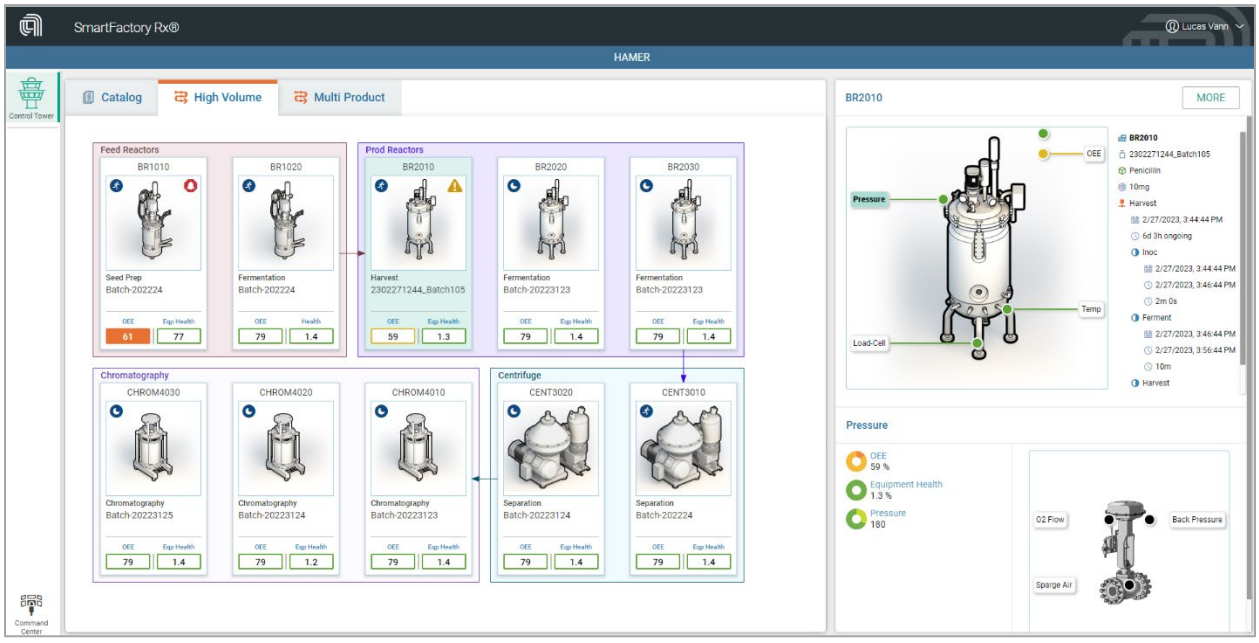

Figure 4. Dashboard example showing high level view of a biomanufacturing process

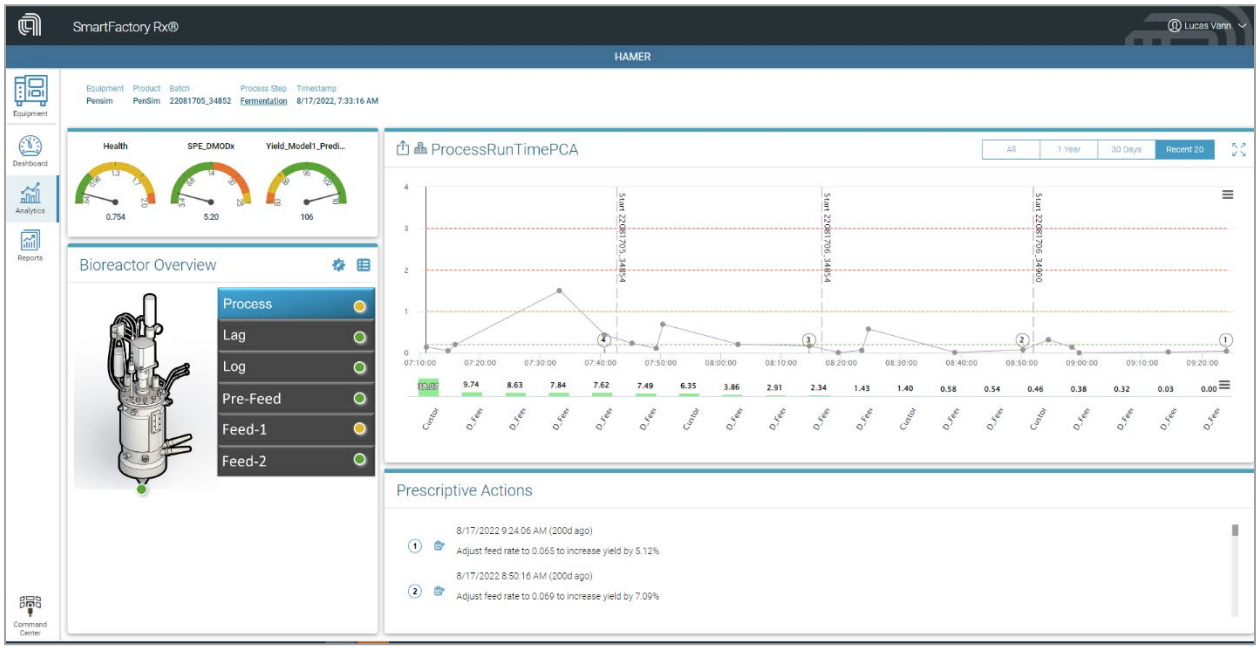

Figure 5. Dashboard example showing prescriptive actions determined by hybrid SGML optimization models

In addition to visualization, automated actions can be applied with the SmartFactory Rx drag-and-drop strategy engine (see Figure 6). These automated actions can be as extensive as:

- Email or text message prescriptive notifications
- Integration to other alerting systems
- Service request / work order generation
- Action plans that initiate workflows for triage of events and dashboard updates for root cause contributors and improvement tracking
- Automated update to maintenance, manufacturing and/or engineering schedules
- Control: feed forward, feed backward and closed-loop

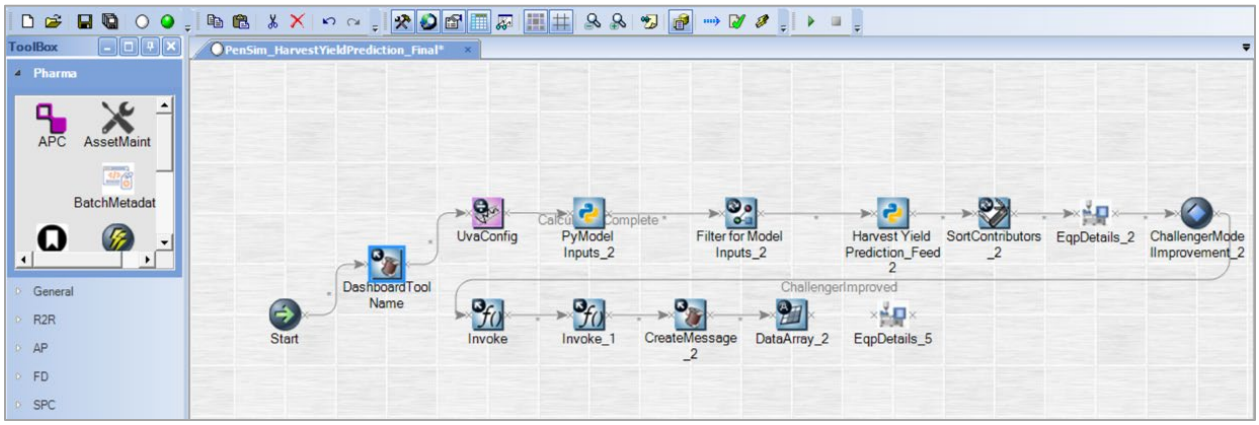

Figure 6. Automated code free orchestration

Summary

Our hybrid SGML model methodology allows for more in-depth monitoring, greater understanding and enables predictive control which is expected to primarily result in increased productivity and consistency from run to run. In general, the current ability in industry to develop and integrate mechanistic models in real-time is extremely limited. This platform demonstrates the capability to integrate multiple data types (advanced sensors, controllers) and data systems (DCS, Historian, CMMS, Building Management System (BMS), Laboratory Information Management System (LIMS)) and provides a platform to consume the data irrespective of the modeling software employed by the end user to generate real-time mechanistic-empirical models. The platform also manages alerts to reduce false alarms, which is a common critical challenge, and is capable of initiating workflows to immediately notify the correct individual with prescriptive actions. All these factors are definitive needs within the industry to advance manufacturing control capabilities. The Expert Unfolding along with feature engineering and multivariate modeling functionality within the software platform fills a current gap within the industry to develop and employ extremely complex mechanistic data transformations as well as fundamental relationship models based on subject matter expertise. Additionally, the mechanistic models can feed into advanced empirical MVA tools to build extremely robust predictive hybrid models. The same platform can be used for both equipment and process models and is extremely suited to predictive maintenance applications so to limit the requirement for multiple solutions and the resulting increased unnecessary maintenance and compromised uptime.

Estimated time to impact:

The estimated time to impact post deployment is almost immediate. Once SmartFactory Rx is installed at the manufacturing site and the equipment templates are populated and deployed to supervise the manufacturing process in real-time, it is used immediately to create knowledge and wisdom from the data and information at the site. As SmartFactory Rx is further leveraged in moving from a stepped process of “information only” to “automated action,” the impact to the industry dramatically increases. It will enable biomanufacturing companies to lower the cost of goods and assist them in getting their products to patients faster. Any GMP manufacturing facility should of course go through the due diligence of validation to ensure they are complying with governing regulations. This will have a time impact dependent on the level of use. However, a number of the top 10 Pharma companies have leveraged SmartFactory Rx to generate high return on investments within the first year of operation.

References:

Biechele, P., Busse, C., Scheper, T., & Reardon, K. (2015). Sensor systems for bioprocess monitoring. Eng. Life Sci. *15*, 469–488.

Chopda, V.R., Gomes, J., and Rathore, A.S. (2016). Bridging the gap between PAT concepts and implementation: An integrated software platform for fermentation. Biotechnol. J. *11*, 164–171.

Graham, L.J. (2016). Leveraging Data Analytics Innovations to Improve Process Outcomes. Biopharm Int. North Olmsted *29*, 18–22.

Hausmann, R., Henkel, M., Hecker, F., and Hitzmann, B. (2017). 25 - Present Status of Automation for Industrial Bioprocesses. In Current Developments in Biotechnology and Bioengineering, C. Larroche, M.Á. Sanromán, G. Du, and A. Pandey, eds. (Elsevier), pp. 725–757.

Izat, N., Yerlikaya, F., and Capan, Y. (2014). A glance on the history of pharmaceutical quality by design. OA Drug Des. Deliv.*2*,1–8.

Kyriakopoulos, S., Ang, K., Lakshmanan, M., Huang, Z., Yoon, S., Gunawan, R., Lee, D. (2018). Kinetic Modeling of Mammalian Cell Culture Bioprocessing: The Quest to Advance Biomanufacturing. Biotechnology Journal. *13*. 1-11.

Langer, E. (2013). The Future of Biopharma. Biopharm Int. North Olmsted *26*, 22–24.

Marx, V. (2013). The Big Challenges of Big Data. Nat. Lond. *498*, 255–260.

Mayalu M and Asada H. (2014). An information-theoretic approach to integrated mechanistic-empirical modeling of cellular response based on intracellular signaling dynamics American Control Conference

Quiroga-Campanoa A., Panoskaltsisa, N., Mantalarisa, A. (2018). Energy-based culture medium design for biomanufacturing optimization: A case study in monoclonal antibody production by GS-NS0 cells. Metabolic Engineering. *47*, 21-30.

Read, E.K., Park, J.T., Shah, R.B., Riley, B.S., Brorson, K.A., and Rathore, A.S. (2010). Process analytical technology (PAT) for biopharmaceutical products: Part I. concepts and applications. Biotechnol. Bioeng. *105*, 276–284.

Rios, M. (2014). Analytics for Modern Bioprocess Development. BioProcess Int. *12*, 1–8.

Sharma N., and Liu Y.A. (2022). A hybrid science-guided machine learning approach for modeling chemical processes: A review. AIChE, Process Systems Engineering, Volume 68, Issue 5, 1-19.

Ündey, C., Ertunç, S., Mistretta, T., and Looze, B. (2010). Applied advanced process analytics in biopharmaceutical manufacturing: Challenges and prospects in real-time monitoring and control. J. Process Control *20*, 1009–1018.

Yahia B., Malphettes L., Heinzle, E. (2021). Predictive macroscopic modeling of cell growth, metabolism and monoclonal antibody production: Case study of a CHO fed-batch production. J Metabolic Engineering *66*, 204-216.

Contact

SmartFactory_Rx@amat.com

appliedsmartfactory.com/pharmaceutical

APPLIED MATERIALS® | Applied SmartFactory®