Computers and Chemical Engineering 179 (2023) 108445

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Smart process analytics for the end-to-end batch manufacturing of monoclonal antibodies

Moo Sun Hong [a,b,1], Fabian Mohr [a,1], Chris D. Castro [c], Benjamin T. Smith [c], Jacqueline M. Wolfrum [d], Stacy L. Springs [d], Anthony J. Sinskey [d,e], Roger A. Hart [c], Tom Mistretta [c], Richard D. Braatz [a,d,*]

[a] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[b] School of Chemical and Biological Engineering, Seoul National University, Seoul, South Korea
[c] Amgen, One Amgen Center Drive, Thousand Oaks, CA, USA
[d] Center for Biomedical Innovation, Massachusetts Institute of Technology, Cambridge, MA, USA
[e] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Process data analytics
Machine learning
Biopharmaceutical manufacturing
Biomanufacturing
Monoclonal antibodies

ABSTRACT

For many modern biopharmaceutical processes, manufacturers develop data-driven models using data analytics/machine learning (DA/ML) methods. The challenge is how to select the best methods for a specific dataset to construct the most accurate and reliable model. This article describes the application of smart process data analytics software to industrial end-to-end biomanufacturing datasets for monoclonal antibody production to automate the determination of the best DA/ML tools for model construction and process understanding. The application demonstrates that smart process data analytics software captures product- and process-specific characteristics for two different monoclonal antibody productions. This study provides tools that can be widely applied to biomanufacturing processes for root cause analysis, prediction, and control.

1. Introduction

Biopharmaceuticals, which are drug products derived from biological organisms for treating or preventing diseases, are continuously growing in terms of both global sales and pipeline due to many advantages such as high specificity and activity (Hong et al., 2018). As the worldwide pharmaceutical sales continue to grow and forecast to top $1 trillion by 2026, biopharmaceuticals will account for 37 % of sales in 2026, up from 30 % in 2020 (Evaluate Pharma, 2021). By 2026, more than half of the 100 top-selling medicines will be biopharmaceuticals, generating 57 % of the sales from this group (Evaluate Pharma, 2021). The percentage of the biopharmaceuticals in the total pipeline also gradually increased for over 20 years, reaching more than 40 % in 2020 (Lloyd, 2021). Monoclonal antibodies (mAbs) are the second largest category of drugs in the current pipeline, which is after the general class of small molecules produced via traditional synthetic chemistry techniques (Lloyd, 2021).

The continued evolution of biologics manufacturing and quality expectations has led to the application of process analytical technology (PAT), which is defined by the Food and Drug Administration (FDA) as "a system for designing, analyzing, and controlling manufacturing through timely measurements (i.e., during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality" (FDA, 2004). Measurements of critical quality attributes (CQAs) through PAT allow deeper understanding and model construction of biopharmaceutical manufacturing processes. Depending on the degree of process understanding, the process models range from data-driven to mechanistic (WHO, 2008).

For many biopharmaceutical manufacturing processes, mechanistic models are not available due to the lack of complete biological process understanding and analytical quantification (Hong et al., 2018; Hong et al., 2020; Narayanan et al., 2020). This situation has required manufacturers to develop data-driven models using data analytics (DA)/machine learning (ML) methods (Hong et al., 2020; Jiang et al., 2017; Steinwandter et al., 2019; Smiatek et al., 2020; Banner et al., 2021;

* Corresponding author: Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.
E-mail address: braatz@mit.edu (R.D. Braatz).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.compchemeng.2023.108445
Received 29 April 2023; Received in revised form 31 August 2023; Accepted 30 September 2023
Available online 2 October 2023

Maruthamuthu et al., 2020). Past studies have already demonstrated that DA/ML methods can construct accurate and reliable models for industrial biomanufacturing processes by identifying and building relationships between critical process parameters (CPPs) and CQAs (Maruthamuthu et al., 2020; Abu-Absi et al., 2010; Rathore et al., 2015; Severson et al., 2018; Severson et al., 2015). These studies have used conventional multivariable statistical approaches in conjunction with contemporary machine learning techniques. Some of the studies incorporated model inputs such as raw material properties and cell culture process variables (Abu-Absi et al., 2010; Rathore et al., 2015). Tools and software packages that leverage DA/ML are rapidly increasing in availability; therefore, the challenge today is how to *select* the best methods and tools for a specific biomanufacturing dataset to ensure that most accurate and reliable model is constructed (Hong et al., 2020; Maruthamuthu et al., 2020). The optimal selection of DA/ML tools requires a substantial level of expertise due to the diverse nature of biomanufacturing data in terms of both quantity and quality. In the absence of this expertise, the inadvertent selection of suboptimal DA/ML tools can result in less accurate predictions and lower quality decisions made based on those predictions.

This challenge motivated the development of a smart process analytics (SPA) software, which automates the selection of methods and construction of models (Severson et al., 2018; Severson et al., 2015; Sun and Braatz, 2021). This article describes the first application of smart process data analytics software to industrial end-to-end biomanufacturing datasets for monoclonal antibody production to determine the best DA/ML tools for model construction and process understanding. The application analyzes the specific data characteristics present in biomanufacturing datasets and proposes software enhancements based on the discovered insights. Section 2 briefly describes the overall structure of smart process data analytics. Section 3 describes the industrial biomanufacturing datasets used in the case studies. Section 4 discusses the application procedure and results through comparison between different methods, followed by the conclusions in Section 5.

2. Smart process analytics (SPA)

Smart process analytics (SPA) is a software which automatically selects DA/ML tools for manufacturing data based on specific characteristics of the data and expert domain knowledge of the process (Hong et al., 2020; Sun and Braatz, 2021). This Python-based software has a user-friendly interface that requests datasets and the modeling objective from the user and then provides the final model with its performance.

Based on the data characteristics and whether the objective is for the model to be interpretable, SPA selects the most suitable DA/ML tool by following a decision tree which can be represented in the form of a triangle (Fig. 1). This data analytics triangle was developed from literature review, theoretical analyses, and case studies to provide modeling techniques that are most suitable for data with the specific characteristics (Sun and Braatz, 2021). In terms of interpretability, the user can specify whether predictive accuracy is the only modeling objective or the model should also be *sparse*. Sparse models, which identify a subset of predictors, are more interpretable and robust when the number of samples is limited. On the other hand, dense models which use all of the predictors produce highly accurate predictions but can be prone to overfitting when the number of samples are limited.

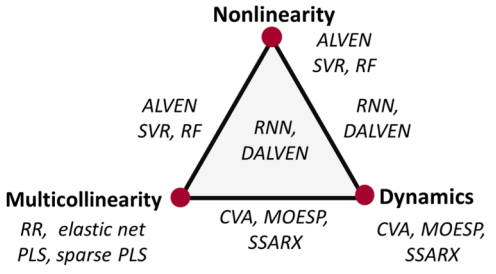

Fig. 1. The data analytics triangle for predictive modeling with a single response variable. The modeling techniques are mapped to three data characteristics. ALVEN = algebraic learning via elastic net; CVA = canonical variate analysis; DALVEN = dynamic ALVEN; MOESP = multivariable output error state space; PLS = partial least squares; RF = random forest regression; RNN = recurrent neural network; RR = ridge regression; SVR = support vector regression. Adapted from Sun and Braatz (2021).

With the given datasets, SPA first assesses the specific data characteristics based on nonlinearity, multicollinearity, and dynamics (Fig. 1). These characteristics violate the assumptions of the Gauss-Markov theorem, which states conditions for which ordinary least squares produces the best linear unbiased estimation. Nonlinearity is judged to be significant if the bilinear correlation test, quadratic test, or maximal correlation analysis show statistically significant nonlinearity (Sun and Braatz, 2021). Multicollinearity occurs in essentially all datasets for the manufacturing of drug substance (Maruthamuthu et al., 2020; Severson et al., 2018). This study does not consider time series data, so dynamics within each batch are not represented in the data. Dynamic batch-to-batch effects, on the other hand, can play a role in batch manufacturing as media lots and resin may be used for multiple batches.

SPA constructs the final model using a rigorous cross-validation procedure for optimal tradeoff between robustness (variance) and accuracy (bias). The datasets are split into training, validation, and testing, where each group of datasets is used to train the model, select hyperparameters, and evaluate the unbiased model error. Depending on the time available for cross validation, the software uses a nested cross-validation procedure to also estimate model stability. This procedure splits the datasets into many different train-test sets for the model construction and evaluation.

All DA/ML tools in the data analytics triangle are well established in the literature, except for algebraic learning via elastic net (ALVEN) and its generalization to dynamic systems, dynamic ALVEN (Sun and Braatz, 2020). ALVEN and dynamic ALVEN combine expert knowledge (biological and chemical nonlinear transformations) with machine learning (sparse regression to deal with a small number of batches) to construct a nonlinear interpretable model. The final model is selected from a large library of informative linear and nonlinear terms through the two-step sparsity-promoting technique to provide interpretability and furthermore physical/chemical insights of the system.

3. Biopharmaceutical manufacturing datasets

Smart process analytics was applied to manufacturing datasets for two monoclonal antibodies.

3.1. Monoclonal antibody #1 (mAb#1)

Chinese hamster ovary (CHO) cells are used to produce the monoclonal antibody. The cell culture process consists of a series of scale-up and expansion steps which generate sufficient cell mass for inoculation of the production bioreactor, additional cell growth and production of the monoclonal antibody, and removal of cellular mass from the bioreactor material using centrifugation and three stages of filtration to obtain the clarified material, Harvested Cell Culture Fluid (HCCF) (Fig. 2a). The downstream purification process starts with acid precipitation of the HCCF and continues through chromatography column 1, low pH hold for viral inactivation, chromatography columns 2 and column 3, viral filtration, and ultrafiltration/diafiltration (UF/DF) to yield the Bulk Drug Substance (BDS) (Fig. 2b).

The critical quality attribute which serves as the model output for this dataset is the antibody's basic peaks percentage, measured at the column 2 pool (Fig. 2b). Basic peaks contain oxidized product-related impurities such as N- and C-terminal variants as well as high molecular weight (HMW) species. The basic peaks percentage may be impacted by various process steps, including operating conditions in the

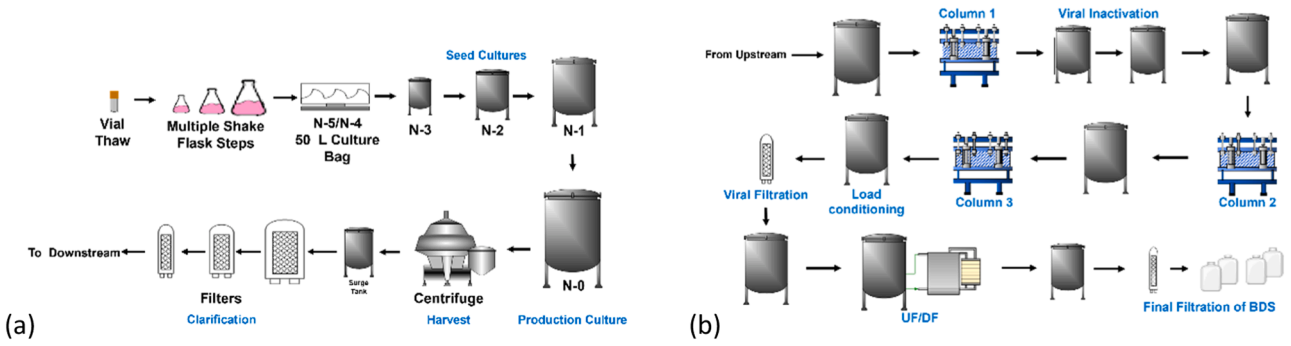

Fig. 2. Process overview of mAb#1 (a) upstream and (b) downstream.

production bioreactor. To model basic peaks percentage, the number of input variables (measurements) is 169 (136 from the bioreactors, 6 from the harvest, 9 from column 1, 10 from viral inactivation, and 8 from column 2). The number of data points (observations) is 77, collected over many manufacturing campaigns.

3.2. Monoclonal antibody #2 (mAb#2)

CHO cells are used to produce mAb#2 in a process similar to mAb#1, albeit with some differences in the downstream unit operations (Fig. 3).

The model output for the mAb#2 dataset is the product's pre-monomer percentage measured at the drug substance pool. The pre-monomer has been identified as a complex of an intact monoclonal antibody covalently bound to a fragment containing a light chain and a variant of the heavy chain. The pre-monomer percentage may be impacted by various process steps, particularly column 2. To model pre-monomer percentage, the number of input variables (measurements) is 120 (89 from the bioreactors, 4 from the harvest, 8 from column 1, 8 from viral inactivation, and 11 from column 2). The number of data points (observations) is 94, collected over many manufacturing campaigns.

4. Results and discussion

4.1. Data interrogation

The first step is to check for linear correlation using the (Pearson) correlation coefficient,

$$\rho_{x_i,y} = \mathrm{corr}(x_i, y) = \frac{\mathrm{cov}(x_i, y)}{\sigma_{x_i}\sigma_y}, \quad i = 1, \ldots, m, \tag{1}$$

where $x = [x_1 \cdots x_m]^{\mathrm{T}} \in \mathbb{R}^m$ is a vector of predictors, y is a univariate response variable, cov is the covariance, and σ is the standard deviation. The correlation coefficient ranges from -1 to 1 with a higher absolute value corresponding to stronger linear correlation. For the mAb#1 and mAb#2 datasets, predictors that have high correlation coefficient with the response variable are mainly from upstream for mAb#1 (Table 1) and downstream for mAb#2 (Table 2). Magnitudes of high correlation coefficients also show that the mAb#1 dataset has more frequent strong linear correlations between the predictors and the response variable.

The next step is to assess nonlinear correlations using the quadratic and the maximal correlation tests. The two tests are used together in SPA for a thorough analysis (Sun and Braatz, 2021); the maximal correlation is presented here since it may be unfamiliar to some readers. The maximal correlation (Rényi, 1959) is defined as

$$\rho^*_{x_i,y} = \sup_{\theta,\phi} \mathrm{corr}(\theta(x_i), \phi(y)), \quad i = 1, \cdots, m, \tag{2}$$

where θ and ϕ run over all real-valued functions with zero mean and finite variances. The maximal correlation ranges from 0 to 1 and includes linear correlations. Consequently, a higher difference of the maximal correlation to the absolute value of the linear correlation coefficient corresponds to stronger nonlinear correlation. Some inputs with high differences between the maximal correlation measure and the absolute linear correlation coefficient indicate nonlinearity in the datasets for both mAb#1 and mAb#2 (Tables 3 and 4, respectively).

Multicollinearity is assessed by the variable inflation factor (VIF) in SPA (Sun and Braatz, 2021), but is not needed for this analysis since the larger number of predictors than samples for both datasets implies high multicollinearity.[2] High intercorrelation between the predictors can be also quantified by the correlation matrix of the predictors (Fig. 4). The number of predictors with correlation greater than 0.9 with each other is 53 and 34 for the mAb#1 and mAb#2 datasets respectively, which is about 30 % of total number of predictors for both datasets. This large number of strong correlations between the predictor variables is an additional indicator of multicollinearity in the dataset.

The data interrogation from SPA indicates the presence of nonlinearity and multicollinearity in both datasets, which SPA indicates that ALVEN, support vector regression (SVR), or random forest regression (RF) should be applied (Fig. 1). ALVEN is a sparse modeling method, which identifies a subset of predictors, whereas SVR and RF are dense modeling methods which use all predictors, including predictors that are not useful for making predictions. The nonlinearity analysis in SPA also identifies nonlinearities associated with input variables whose inclusion in the model would not significantly increase its predictive accuracy. For example, a single input variable that shows a significant nonlinear correlation with the output will result in the nonlinearity analysis recommending application of nonlinear modeling methods, although a highly accurate linear model may be constructed using only the input variables that have a linear relationship to the model output. To evaluate the effectiveness of SPA in selecting the best modeling methods for these biopharmaceutical manufacturing datasets, as well as evaluating the nonlinearity analysis more broadly for datasets in which nonlinear input-output relationships may occur but may not significantly improve the predictive accuracy, we compare the methods selected by SPA to linear multicollinear modeling methods that have been demonstrated to produce very high predictive accuracy for such datasets (Sun and Braatz, 2021). Namely, we compare to Partial Least Squares (PLS), which is the most widely used dense method for building models for linear multicollinear data, and sparse PLS (SPLS) and Elastic Net (EN) which are sparse modeling methods. Sparse methods such as ALVEN, EN, and SPLS are more interpretable and robust when the number of samples is limited (Sun and Braatz, 2021).

[2] Data from biopharmaceutical manufacturing processes are multicollinear, even when there are more samples than predictors, as the variables are related by governing laws such as conservation of mass, conservation of energy, and kinetic reaction equations (Ferreira et al., 2018).

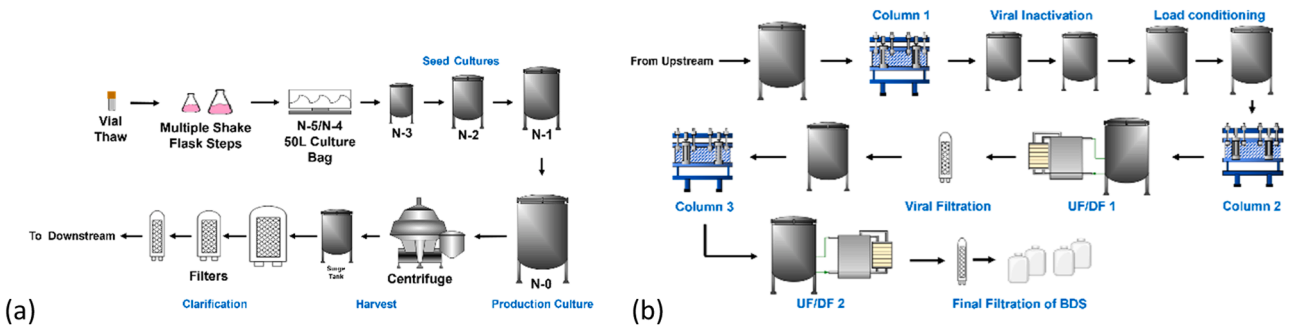

Fig. 3. Process overview of mAb#2 (a) upstream and (b) downstream.

Table 1
Pearson correlation coefficients for mAb#1 data ($|\rho| > 0.6$).

ρ	Unit Operation	Parameter
0.7565	N Production Bioreactor	Product Mass
0.7064	N Production Bioreactor	Volume – Final Day
0.6990	N Production Bioreactor	Cumulative Antifoam
0.6943	Harvest	Product Mass
0.6614	N Production Bioreactor	Cumulative Acid
0.6045	N-1 Bioreactor	Viability – Final Day
0.6044	Column 1	Product Mass

Table 2
Pearson correlation coefficients for mAb#2 data ($|\rho| > 0.4$).

ρ	Unit Operation	Parameter
−0.7594	Column 2	Pool Volume
0.5521	Column 2	Resin Cycle Number
0.4841	Column 2	Product HMW%
−0.4221	Column 1	Resin Cycle Number

Table 3
Maximal correlation coefficients for mAb#1 data ($\rho^* > 0.6, |\rho| < 0.2$).

$\rho^* - \lvert\rho\rvert$	ρ	$\rho^* - \lvert\rho\rvert$	Unit Operation	Parameter
0.7480	0.0018	0.7462	N Production Bioreactor	Glucose – Day 8
0.8145	0.1688	0.6458	Harvest	Discharge Interval
0.6279	0.0526	0.5753	Column 1	Pool Volume
0.6279	0.0526	0.5753	Viral Inactivation	Pool Volume
0.6643	0.1697	0.4946	Viral Inactivation	Mean Temperature

Table 4
Maximal correlation coefficients for mAb#2 data ($\rho^* > 0.6, |\rho| < 0.2$).

$\rho^* - \lvert\rho\rvert$	ρ	$\rho^* - \lvert\rho\rvert$	Unit Operation	Parameter
0.8954	0.0712	0.8242	Column 2	Product Monomer
0.6332	0.1002	0.5330	N Production Bioreactor	Total Cell Density – Day 0
0.6236	0.0987	0.5249	Harvest	Product Mass
0.6091	0.1131	0.4960	Column 2	Host Cell Protein

4.2. Model construction

The model construction procedures were conducted for both datasets for the aforementioned 6 models: ALVEN, SVR, RF, PLS, SPLS, and EN. For the ALVEN models, the complexity of the nonlinear transformation was limited to degree 1 to only consider basic transformations for simplicity and interpretability of the model (Sun and Braatz, 2020). Higher degrees allow for the multiplication of different features which might be predictive but are not easily interpreted.[3] Nested cross-validation with repeated k-fold cross-validation for the inner loop is implemented for model construction. Five-fold cross-validation was repeated 10 times for the inner loop, and the outer loop is repeated for 15 train-test splits to ensure a comparison of the models that is statistically reliable. There is an inherent tradeoff of error metric mean and variance between the size of training and test data. Commonly used train-test ratios are 80–20, 70–30 or 67–33. To guarantee an accurate estimation of the error variances in the smaller testing data set, it is recommended to have at least 20 to 30 samples in this set (Murphy, 2012). Based on this, we use a 70–30 train-test split for the outer loop. The model construction procedures were repeated for the logarithm of the response variable to consider broader forms of nonlinear models (Sun and Braatz, 2021; Severson et al., 2019).

The results for the mAb#1 dataset are shown in Table 5 and Fig. 5. For this dataset, the linear models (PLS, SPLS, and EN) are overall better than nonlinear models (ALVEN, SVR, and RF) in terms of model accuracy of the testing dataset. The linear models show consistent means of training and test errors, while the nonlinear models perform much worse for testing dataset indicating significant overfitting by the nonlinear methods. This result shows that nonlinearity is not significant enough in the mAb#1 dataset to justify the usage of nonlinear models like ALVEN. PLS is the best performing algorithm for prediction without interpretability and SPLS is the best candidate considering interpretability for root cause analysis. Comparing the models utilizing a log transformation or no log transformation, both models have similar accuracies; however, the models for the logarithm of output are slightly better than the models for the output without transformation.

Among linear models, coefficients of the SPLS model provide interpretable information of the model (Table 6). Predictors included in a greater number of splits with larger coefficients are likely to be included in the final model formulation depending on the modeling purpose. As seen from the data interrogation (Table 1), predictors from the upstream with high correlation coefficients with the response variable are frequently included into the SPLS model.

The impact of the N Production Bioreactor Mass on the basic peak percentage is consistent with the cell biology. This mass is the product of two parameters:

(1) The N Production Bioreactor Volume. The CHO cells tend to experience the same conditions (e.g., temperature, pH, culture duration, nutrient concentrations) regardless of production bioreactor volume. When the bioreactor volume increases, such as due to process improvements, the rate at which gas is sparged to oxygenate cells changes. At the higher gas flow rates used for

[3] We also constructed higher order ALVEN models, but the models did not have higher prediction accuracy and were more prone to overfitting.

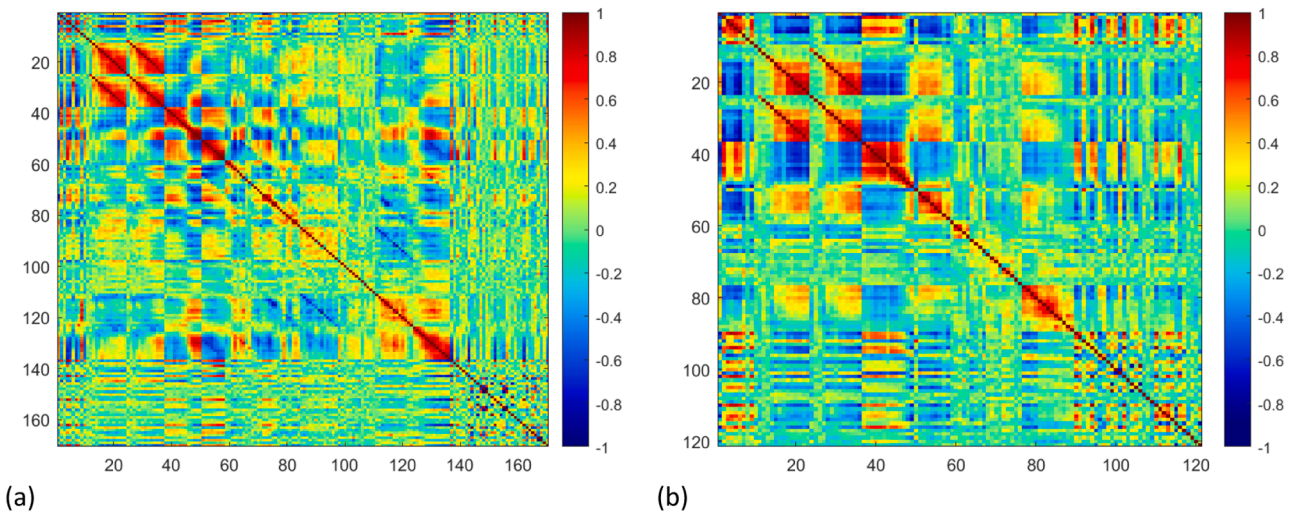

Fig. 4. Graphs of correlation matrices of the predictors for (a) mAb#1 and (b) mAb#2 data. Many predictors are highly correlated with each other.

Table 5
Model prediction results for mAb#1 data.

Model	Output without Transformation				Logarithm of Output			
	Train		Test		Train		Test	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ALVEN	0.345	0.039	0.457	0.075	0.322	0.071	0.453	0.106
SVR	0.245	0.014	0.485	0.074	0.235	0.014	0.467	0.093
RF	0.330	0.029	0.436	0.069	0.332	0.028	0.447	0.118
PLS	0.457	0.017	0.400	0.065	0.441	0.029	0.408	0.078
SPLS	0.470	0.040	0.444	0.084	0.465	0.032	0.421	0.096
EN	0.516	0.056	0.492	0.083	0.507	0.061	0.477	0.105

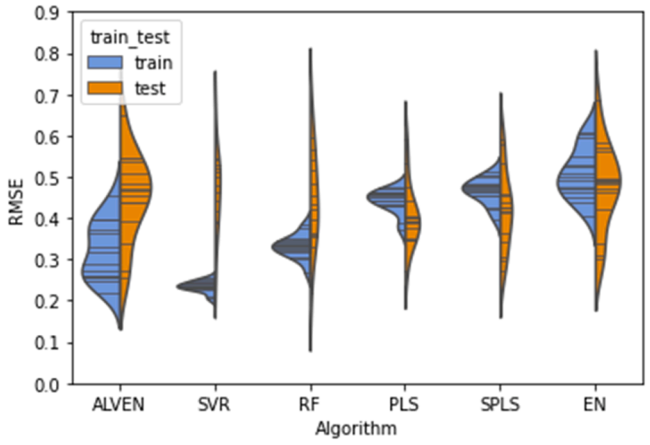

Fig. 5. Model prediction results for mAb#1 data with logarithm of output.

Table 6
Coefficients of SPLS model for mAb#1 data with logarithm of output (Means are computed for nonzero coefficients).

# Splits	Mean	Unit Operation	Parameter
15	0.5803	N Production Bioreactor	Product Mass
3	0.3460	Harvest	Product Mass
2	0.3914	N Production Bioreactor	Volume – Final Day
2	0.3813	N Production Bioreactor	Cumulative Acid

larger culture volumes, the gas entrance velocity increases, subjecting the CHO cells to greater shear stress. This stress can adversely affect growth, product expression, and even cell viability. Cells may produce and secrete glycoproteins of different product quality when subjected to greater stresses, impacting parameters like the basic peak percentage.

(2) The N Production Bioreactor Titer. Differences in bioreactor titer can be associated with changes in product quality. For example, if an increase in bioreactor titer is due to an increase in specific productivity, the CHO cells have less time to create and secrete individual product molecules. The product molecules travel faster through modifying cellular organelles like the Golgi apparatus, which can affect product quality parameters.

Table 6 also lists three other parameters, likely because they have direct or indirect correlations to the N Production Bioreactor Mass parameter. For example, both the N Production Bioreactor Volume – Final Day and Harvest Mass directly correlate with N Production Bioreactor Mass. Furthermore, the amounts of Cumulative Acid tend to positively correlate with N Production Bioreactor Volume – Final Day, as larger cultures require more acid and more antifoam.

The results for the mAb#2 dataset are shown in Table 7 and Fig. 6. For this dataset, the model accuracy is slightly better for the nonlinear models compared to the linear models except for SVR. SVR models have the most extreme difference between training and testing for both datasets, indicating that the model is strongly overfitted. Nonlinearity is more important for mAb#2 compared to mAb#1 and considering the interpretability, ALVEN is the best candidate. For prediction, RF is the best performing algorithm as interpretability is not necessary for models whose sole purpose is prediction. Models constructed for the output and the logarithm of output again have similar accuracies, but the ALVEN

Table 7
Model prediction results for mAb#2 data.

Model	Output without Transformation				Logarithm of Output			
	Train		Test		Train		Test	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
ALVEN	0.0309	0.0075	0.0413	0.0076	0.0311	0.0076	0.0421	0.0070
SVR	0.0192	0.0058	0.0476	0.0086	0.0211	0.0067	0.0470	0.0089
RF	0.0298	0.0022	0.0403	0.0042	0.0297	0.0025	0.0407	0.0040
PLS	0.0325	0.0056	0.0460	0.0074	0.0323	0.0056	0.0459	0.0073
SPLS	0.0427	0.0046	0.0454	0.0061	0.0422	0.0045	0.0456	0.0059
EN	0.0349	0.0049	0.0411	0.0064	0.0351	0.0049	0.0410	0.0064

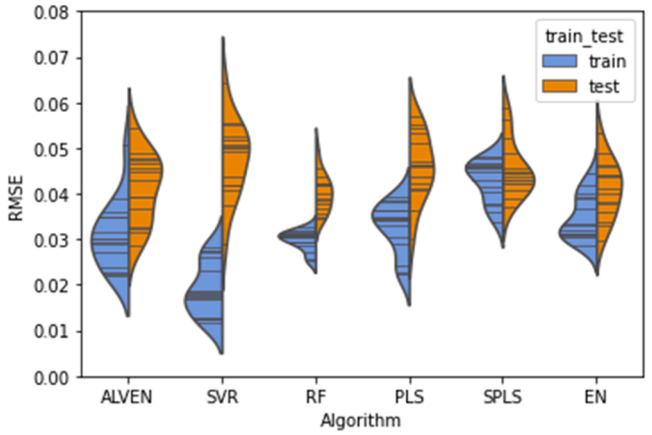

Fig. 6. Model prediction results for mAb#2 data.

model for the output without transformation is slightly better than that for the logarithm transformed output.

Predictors that are included in the ALVEN model are mostly from downstream as identified in the data interrogation (Table 8). Some of the nonlinear predictors included in the ALVEN model are those that are detected in the nonlinearity test or highly correlated to those predictors. This is the reason why the nonlinearity became important in the mAb#2 dataset but not in the mAb#1 dataset.

The pre-monomer impurity is created in the N Production Bioreactor, and the pre-monomer percentage is affected by variations in the nutrient concentrations. During the manufacturing process, a key activity is the Day 4 nutrient feed, where nutrient variation at commercial scale has been theorized to cause significant variations in growth. In turn, as cell growth varies, so can the product quality aspects of the product molecule. Osmolality – Day 4 may serve as an indicator that for some production lots, additional feed (or additional, specific nutrients) may aid cell nutrition and in turn reduce pre-monomer impurity levels.

Pre-monomer levels are also impacted by purification steps, where process variability can also result in differences in the amount of impurities being removed. For example, as Column 2 pool volume and therefore product mass increases, there are relatively fewer resin binding sites on the Column 2 resin. As such, the resin may preferentially retain the smaller monomer, while more pre-monomer species flow through the column. For Column 1, when more product monomer is retained, more pre-monomer species flow through the column.

Table 8
Coefficients of ALVEN model for mAb#2 data that are included at least in 8 splits (means are computed for nonzero coefficients).

# Splits	Mean	Term	Unit Operation	Parameter
12	−0.4062	linear	Column 2	Pool Volume
11	−0.1079	linear	Column 1	Product Monomer%
10	−0.1088	log	Harvest	Yield
9	−0.1421	sqrt	Column 2	Pool Volume
8	−0.3405	log	Column 2	Pool Volume
8	−0.1512	linear	N Production Bioreactor	Osmolality – Day 4
8	−0.0652	sqrt	N Production Bioreactor	Osmolality – Day 4

4.3. Modeling use cases

For biopharmaceutical manufacturing, model decisions such as choice of inputs and model selection are likely to depend on the use case, whether root cause analysis, prediction, or control.

For prediction, model selection is focused on high predictive accuracy and may be able to use dense or non-interpretable models. The input variables may include all potentially relevant parameters, including operationally dependent features.

For root cause analysis, model interpretability is a necessity. A similarly broad set of input variables may be used. Note that, for data sets with multicollinearity, the model is picking the model parameters that are best correlated with the response and do not necessarily represent root-cause parameters. Interpreting the root cause from among the model parameters still requires subject matter expertise about the process.

Lastly, modeling for control, whether active or offline, requires a specific model that relates the response to be controlled to what input parameters can be operationally changed. For example, models introduced in Section 4.2 would not be appropriate for control as operationally dependent parameters were also included in the input variables. The input variables must be pared down to reflect only the parameters that can be controlled. Model selection will be determined based on a combination of predictivity and interpretability requirements.

5. Conclusion

Smart process data analytics software is applied to industrial end-to-end biomanufacturing datasets for two different monoclonal antibody (mAb) products to construct models based on the best data analytics/machine learning tools. The dataset for mAb#1 is modeled with Partial Least Squares (PLS) for the logarithm of the output, and predictors from the upstream are mainly used for the model prediction. On the other hand, the model for the dataset for mAb#2 is constructed with Algebraic Learning Via Elastic Net (ALVEN) for the output without transformation, and predictors for the model prediction are heavily chosen from the downstream. As shown from the dataset for mAb#1, data with nonlinearity may be best described by a linear model if the predictors indicating the nonlinearity are not included in the constructed model. The capability of smart process data analytics software to capture product- and process-specific characteristics enables wide application of the software to biomanufacturing processes.

CRediT authorship contribution statement

Moo Sun Hong: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Fabian Mohr:** Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing –

review & editing, Visualization. **Chris D. Castro:** Conceptualization, Validation, Data curation, Writing – original draft, Writing – review & editing. **Benjamin T. Smith:** Conceptualization, Validation, Data curation, Writing – original draft, Writing – review & editing. **Jacqueline M. Wolfrum:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Stacy L. Springs:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Anthony J. Sinskey:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Roger A. Hart:** Conceptualization, Resources, Writing – review & editing, Funding acquisition. **Tom Mistretta:** Conceptualization, Resources, Writing – review & editing. **Richard D. Braatz:** Conceptualization, Methodology, Resources, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This study is supported by the U.S. Food and Drug Administration, Contract No. 75F40121C00090. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the financial sponsor.

References

Abu-Absi, S.F., Yang, L., Thompson, P., Jiang, C., Kandula, S., Schilling, B., Shukla, A.A., 2010. Defining process design space for monoclonal antibody cell culture. Biotechnol. Bioeng. 106 (6), 894–905.

Banner, M., Alosert, H., Spencer, C., Cheeks, M., Farid, S.S., Thomas, M., Goldrick, S., 2021. A decade in review: use of data analytics within the biopharmaceutical sector. Curr. Opin. Chem. Eng. 34, 100758.

Evaluate Pharma,, 2021. World Preview 2021. Outlook to 2026. Evaluate, London, United Kingdom.

FDA, 2004. Guidance For Industry: PAT – A Framework For Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. U.S. Food and Drug Administration, Rockville, Maryland.

Ferreira, A.P., Menezes, J.C., Tobyn, M., 2018. Multivariate Analysis in the Pharmaceutical Industry. Academic Press, London, UK.

Hong, M.S., Severson, K.A., Jiang, M., Lu, A.E., Love, J.C., Braatz, R.D., 2018. Challenges and opportunities of biopharmaceutical manufacturing control. Comput. Chem. Eng. 110, 106–114.

Hong, M.S., Sun, W., Lu, A.E., Braatz, R.D., 2020. Process analytical technology and digital biomanufacturing of monoclonal antibodies. Am. Pharm. Rev. 23 (6), 122–125.

Jiang, M., Severson, K.A., Love, J.C., Madden, H., Swann, P., Zang, L., Braatz, R.D., 2017. Opportunities and challenges of real-time release testing in biopharmaceutical manufacturing. Biotechnol. Bioeng. 114, 2445–2456.

Lloyd, I., 2021. Pharma R&D Annual Review 2021. Informa, London, United Kingdom.

Maruthamuthu, M.K., Rudge, S.R., Ardekani, A.M., Ladisch, M.R., Verma, M.S., 2020. Process analytical technologies and data analytics for the manufacture of monoclonal antibodies. Trends Biotechnol. 38 (10), 1169–1186.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, Massachusetts.

Narayanan, H., Luna, M.F., von Stosch, M., Bournazou, M.N.C., Polotti, G., Morbidelli, M., Butté, A., Sokolov, M., 2020. Bioprocessing in the digital age: the role of process models. Biotechnol. J. 15 (1), 1900172.

Rényi, A., 1959. On measures of dependence. Acta Math. Hung. 10 (3–4), 441–451.

Rathore, A.S., Singh, S.K, Pathak, M., Read, E.K., Brorson, K.A., Agarabi, C.D., Khan, M., 2015. Fermentanomics: relating quality attributes of a monoclonal antibody to cell culture process variables and raw materials using multivariate data analysis. Biotechnol. Prog. 31 (6), 1586–1599.

Severson, K., VanAntwerp, J.G., Natarajan, V., Antoniou, C., Thömmes, J., Braatz, R.D., 2015. Elastic net with Monte Carlo sampling for data-based modeling in biopharmaceutical manufacturing facilities. Comput. Chem. Eng. 80, 30–36.

Severson, K., VanAntwerp, J.G., Natarajan, V., Antoniou, C., Thömmes, J., Braatz, R.D., 2018. A systematic approach to process data analytics in pharmaceutical manufacturing: the data analytics triangle and its application to the manufacturing of a monoclonal antibody. Multivariate Analysis in the Pharmaceutical Industry. Academic Press, London, UK, pp. 295–312.

Severson, K.A., Attia, P.M., Norman, J., Perkins, N., Jiang, B., Yang, Z., Chen, M.H., Aykol, M., Herring, P.K., Fraggedakis, D., Bazant, M.Z., Harris, S.J., Chueh, W.C., Braatz, R.D., 2019. Data-driven prediction of battery cycle life before capacity degradation. Nat. Energy 4 (5), 383–391.

Smiatek, J., Jung, A., Bluhmki, E., 2020. Towards a digital bioprocess replica: computational approaches in biopharmaceutical development and manufacturing. Trends Biotechnol. 38 (10), 1141–1153.

Steinwandter, V., Borchert, D., Herwig, C., 2019. Data science tools and applications on the way to Pharma 4.0. Drug Discov. Today 24 (9), 1795–1805.

Sun, W., Braatz, R.D., 2020. ALVEN: algebraic learning via elastic net for static and dynamic nonlinear model identification. Comput. Chem. Eng. 143, 107103.

Sun, W., Braatz, R.D., 2021. Smart process analytics for predictive modeling. Comput. Chem. Eng. 144, 107134.

WHO, 2008. Basic Principles of GMP. World Health Organization, Geneva, Switzerland.