Optimizing Protein Titer Production Using Animal Cells: Predictive Modeling and Recommendations for Enhanced Yield

Abstract ID: 7371

Leave space for author names and affiliations to ensure that the limit of 6 pages is not exceeded. Author names and affiliations to be added before final submission, but not during initial blind review.

Abstract

Current biomanufacturing processes rely heavily on human expertise, struggling to adapt to the growing complexity of bioprocessing. Decision support tools based on machine learning models play a vital role in optimizing the production and timely detection of anomalies in the production line and can save thousands of dollars and reduce quality control costs. We have performed a thorough analysis on a pharmaceutical dataset to identify significant variables that can affect protein titer production. Based on this analysis, we develop machine learning-based models for production of the protein titer in the biopharmaceutical/biotechnology-based manufacturing industry. We envision the practical applications of our models in other biopharmaceutical and biotechnology industries, leading to increased productivity and cost-effectiveness in protein production within animal cells.

Keywords

Biomanufacturing processes, machine learning, protein production, predictive analytics

# Introduction

Recent advancements in biomanufacturing have revolutionized the traditional pharmaceutical industry. A wide range of advanced treatments and medications are now available through biomanufacturing processes, targeting various diseases such as autoimmune disorders, infectious diseases, cancer, and rare genetic disorders, among others [1] Unlike traditional pharmaceutical processes, biomanufacturing utilizes “living” host organisms, such as bacteria, yeast, viruses, or mammalian cells in the production process to produce number of products varying from several proteins to antigens [1]. Of high importance among pharmaceutical products are recombinant proteins, which are frequently used in human healthcare for therapeutics, vaccines, and diagnostic reagents as well as in biochemical analysis[2], [3]. Several types of mammalian cell lines are used for protein production, with Human embryonic kidney and Chinese Hamster Ovary (CHO) being the most frequently employed living organisms [4].

Industrial production of recombinant proteins requires efficient bioprocessing strategies along with strict process monitoring to achieve maximum protein yield [5]. However, despite the recent advances (e.g., high-throughput devices for protein purification [6]), biomanufacturing of recombinant proteins remains a complex and intricate production environment with a heavy reliance on human expertise, particularly for processes monitoring. Machine learning (ML)-based decision aid systems can help eliminate the errors and bias that stem from human-based process monitoring, thus enabling faster and more efficient anomaly detection with potential savings in quality control expenses. Several studies report the success of ML models in improving the process monitoring practices, but the studies on developing ML-based predictive models for recombinant protein biomanufacturing remain extremely limited [7], [8], [9].

In this study, we aim at developing ML-based analytical models to predict the yield of recombinant protein, referred to as “protein titer,” from mammalian cells based on environmental factors, feed strategies, and cell line characteristics. When integrated to the quality control system, such models can aid in early detection of process anomalies, allowing timely corrective actions to turn the faulty process under control. To achieve this aim, we utilize data from Cytovance Biologics, a biomanufacturing firm in Oklahoma City, US. The objectives of this study are twofold, (1) developing ML models using the given dataset designed for the rapid and cost-effective analysis of protein titer production, accurately reflecting titer dynamics; and (2) conducting an in-depth analysis of the ML models to dissect the relationships between input variables and the output variable (titer production). We will identify the variables that yields the highest concentration of recombinant proteins.

# Model Development

## Data

## The dataset includes various experiments conducted by Cytovance, each represented as a record that indicates a unique combination of input variables to assess their impact on a titer production level (n=2,076). All experiment spans 14 to 17 production days, starting data collection from day 3 onwards. Our model utilizes 12 input variables to predict the protein titer (measured) as the primary outcome of interest. The continuous input variables considered are “vessel volume,” where the vessel is the container used for growing cultures, “production day,” “dissolved oxygen (DO),” which is the amount of oxygen (%) present during the experiment, “pH setpoint” which measures the acidity of the environment, “temperature,” “target cell seeding density,” which is the number of cells calculated to start the experiment, “feed added,” and “Glucose Trigger Limit,” which is the lowest allowable concentration of glucose in the experiment. The categorical/binary input variables are “media type (A, B),” which is the type of commercial media used, “feed type,” “CHO strain type (KC, S),” and “supplement,” which is the additional supplement added that is not included in the media or the feed. Table 1 provides the details of the dataset along with the frequency of missing values. The dataset further includes, for each record, 14 outcomes (in addition to the target variable), which are utilized for imputation of missing values in the primary outcome (Table 2).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1. Descriptive statistics of input variables | | | | | | |
| **Variable Name** | **Number of Unique Values** | **Missing  (%)** | **Mean** | **Variance** | **Median** | **Correlation with Titer** |
| Production Day | 15 | 4% | 6.01 | 16.47 | 6 | 0.11 |
| Vessel Volume (L) | 3 | 0% | 0.17 | 0.39 | 0.01 | -0.14 |
| DO | 5 | 26% | 0.42 | 0.01 | 0.5 | 0.21 |
| pH Setpoint | 2 | 27% | 7.02 | 0.00 | 7 | 0.16 |
| Temperature | 3 | 39% | 35.80 | 4.42 | 37 | 0.13 |
| Target Cell Seeding Density (cells/mL) | 7 | 4% | 1.1E+06 | 1.2E+12 | 5.0E+05 | 0.30 |
| Feed Added (%) | 7 | 11% | 1.58 | 2.05 | 2 | 0.02 |
| Glucose Trigger Limit (g/L) | 3 | 5% | 3.74 | 0.47 | 3.5 | 0.20 |
| Supplement | 9 | 4% | - | - | - | - |
| Media Type | 2 | 0% | - | - | - | - |
| Feed Type | 3 | 0% | - | - | - | - |
| Strain | 2 | 0% | - | - | - | - |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 2. Descriptive statistics of the target variable and additional output variables | | | | | | |
| **Variable Name** | **Number of Unique Values** | **Missing  (%)** | **Mean** | **Variance** | **Median** | **Correlation with Titer** |
| Titer by Octet (mg/L) | 675 | 67% | 1988.13 | 4.11E+06 | 1437.37 | 1.00 |
| Viable Cell Density (cells/mL) | 1,369 | 8% | 8.28E+06 | 2.32E+13 | 8.77E+06 | 0.90 |
| Cell Viability (%) | 310 | 7% | 93.26 | 136.70 | 97.60 | 0.76 |
| Average Cell Diameter (µM) | 724 | 10% | 16.14 | 5.54 | 15.97 | 0.72 |
| pH | 772 | 20% | 7.13 | 0.06 | 7.10 | 0.86 |
| Glutamine (mM) | 458 | 62% | 3.63 | 8.30 | 2.86 | -0.44 |
| Glutamate (mM) | 731 | 33% | 6.21 | 10.49 | 5.48 | 0.18 |
| Glucose (g/L) | 626 | 15% | 4.91 | 5.13 | 4.73 | -0.67 |
| Lactate (g/L) | 292 | 18% | 1.13 | 0.63 | 1.04 | -0.82 |
| Ammonium (mM) | 836 | 21% | 5.59 | 17.21 | 4.59 | -0.67 |
| Sodium (mM) | 836 | 20% | 123.79 | 1,710.55 | 108.20 | -0.82 |
| Potassium (mM) | 569 | 20% | 3.99 | 3.10 | 3.84 | -0.87 |
| Calcium (mM) | 18 | 19% | 0.09 | 0.00 | 0.10 | -0.89 |
| Osmolality (mOsm/kg) | 373 | 59% | 417.30 | 23,970.48 | 361.00 | -0.89 |

## Pre-processing

## Upon initial cleaning of the records, we dropped the number of records to *n*=1,636 observations. We use “One-Hot” encoding to transform categorical variables into numerical, followed by data normalization, through mean-centering and scaling, to mitigate the effect of high variance in scale among input data (e.g., between target cell seeding density and pH setpoint). Next, we employ data imputation techniques to treat the missing values in the input variables as well as the target variable. In particular, the titer level data suffers from many missing values (65%). However, our correlation analysis indicates a high correlation between the target variable and the output variable, “cell viable density” (90%). In contrast, the cell viable density variable suffers from far less missing values (8%); hence, we use imputation to recover missing observations of titer based on the cell viable density.

We employ Multiple Imputation by Chained Equations (MICE) [10] method to impute the missing titer data, leveraging the predictive power of the interrelated outcome variables. MICE leverages ML techniques for imputation, and investigate the performance of our imputation using random sampling, least absolute shrinkage and selection operator (LASSO), classification and regression trees (CART), and random forest (RF) regression as the base ML models for MICE. Our choice of using MICE for data imputation is further motivated by the observation that MICE can recover the missing titer levels as non-zero values, particularly in the initial days of the experiments, which would be otherwise “considered” as zero if, for example, time series-based methods were to be used for imputation.

## Model

In order to avoid the so-called “data-leakage” from the training set to the test set, we start with splitting the dataset in training (80%, *n*=328) and test sets (20%, *n*=1,308) prior to any data imputation and hyperparameter tuning. We then apply MICE-based imputation models to the training set using random sampling, RF regression, LASSO, and CART, and compute the value of Root Mean Square Error (RMSE) as a measure of accuracy (see Section 3 for the results). The best data imputation model is then selected as the model that achieves the least RMSE value. We then apply the selected data imputation model to the training set and test sets.

To capture both linear and non-linear dynamics, our strategy includes two models designed for linear relationships (a linear model and an elastic net), two models tailored for non-linear relationships (random forest regression and gradient boosted regression trees (GBRT)), and one model, Multivariate Adaptive Regression Splines (MARS), that serves as a middle ground. This diversified selection ensures a comprehensive analysis across different data behaviors, balancing interpretability with the ability to capture complex, non-linear patterns.

We note that each machine learning model requires specific tuning of its relevant hyperparameters to ensure optimal generalization from the training set to the test data. For this purpose, we employ grid search to evaluate all possible values of hyperparameters in the tuning process. This approach is designed to fine-tune the models for the best possible performance on unseen (test) data. Upon computing optimal values of hyperparameters, we train each model on the entire training set and compare the values of RMSE on the test dataset to select the best ML model. The details of each model, including tuning hyperparameters and their tested values, are outlined in Table 3. All models are implemented in R 4.3.1. We utilize “ranger” and “caret” packages to implement random forest and elastic regression, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3. Machine learning models used and their relevant hyperparameters | | | |
| **Model** | **Hyperparameter** | **Description** | **Values Range** |
| Elastic Net | Alpha | Weight of LASSO/Ridge penalties | 0.0, 0.1, 0.2, ..., 1.0 |
|  | Lambda | Regularization strength | 0.0, 0.1, 0.2, ..., 1.0 |
| Random Forest | MTry | # of variables to possibly split | 1, 4, 7 |
|  | Split Rule | How the tree should determine a split | variance, maxstat, extratrees |
|  | Min. Node Size | Minimum # observations at leaf node | 1, 5, 10, 25 |
| MARS | Degree | Maximum # of feature interactions | 1, 2, 3 |
|  | # Prune | # of terms retained in final model | 2, 5, 7, 10 |
| GBRT | Interaction Depth | # of splits in the trees | 1, 5, 9 |
|  | # Trees | # of trees to fit/create | 50, 100, 150, ..., 500 |
|  | Shrinkage | Learning rate | 0.001, 0.01, 0.1 |
|  | Min. Node Size | Minimum # observations at leaf node | 5, 10, 20 |

# Results and Discussion

Table 4 shows the performance of various ML models within the MICE data imputation technique. Accordingly, it reveals that the choice of base imputation method significantly impacts RMSE. We have repeated training and test 50 times and calculated the average RMSE. Among the ML base methods, MICE with LASSO results in lower average test RMSE. A Tukey comparison of average RMSE reveal that the MICE with LASSO significantly outperforms (*p* < 0.05) other methods. We also note that the MICE approach is able to recover the missing titer levels (as non-zero values), particularly in the initial days of the experiments, which would be otherwise “considered” as zero if time series-based methods were to be used for imputation. MICE approach leverages other output variables to estimate these missing titer values, providing a potentially more accurate reflection of early production stages. We specifically employ cell viable density for imputation due to our correlation analysis revealing a strong correlation (90%) between the target variable and the output variable, cell viable density. Table 5 shows the optimized values of hyperparameters,

After using MICE with LASSO to impute missing values, we compare the use of various machine learning models for titer prediction in terms of RMSE. The methods are compared and ranked from the highest to the lowest performing models in Table 6. Figure 1 shows box plots for average RMSE measure, which have been obtained over 50 iterations on the data for each algorithm. According to Table 6 and Figure 1, a linear regression model trained on data imputed with MICE and LASSO outperforms other machine learning models followed by elastic net. A pairwise comparison indicates that no significant difference (*p* < 0.05) is observed between linear modeling and elastic net.

|  |  |  |
| --- | --- | --- |
| Table 4. The performance of various base ML models within the MICE data imputation | | |
|  |
| **Rank** | **Imputation Method** | **Test RMSE** |  |
| 1 | MICE with LASSO | **1127** |  |
| 2 | MICE with Random Forest | 1155 |  |
| 3 | MICE with CART | 1181 |  |
| 5 | MICE with Random Sampling | 1202 |  |

|  |  |  |
| --- | --- | --- |
| Table 5. Optimal values of hyperparameters | | |
| **Model** | **Hyperparameter** | **Value** |
| Elastic Net | Alpha | 0.495 |
|  | Lambda | 0.955 |
| Random Forest | MTry | 7 |
|  | Min. Node Size | 7.6 |
|  | Split Rule | Variance |
| MARS | # Prune | 10 |
|  | Degree | 2.55 |
| GBRT | # Trees | 427.5 |
|  | Interaction Depth | 8.8 |
|  | Shrinkage | 0.0235 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 6. Comparative results of methods | | | |
| **Rank** | **Regression Model** | **Imputation Method** | **Test RMSE** |
| 1 | Linear Model | MICE with LASSO | **1117** |
| 2 | Elastic Net | MICE with LASSO | 1118 |
| 3 | MARS | MICE with LASSO | 1123 |
| 4 | Random Forest | MICE with LASSO | 1128 |
| 5 | GBRT | MICE with LASSO | 1147 |

A graph with a diagram

Description automatically generated

Figure 11. Performance comparison of employed ML-based models after using MICE with LASSO

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 7. Variable importance ranks by each model | | | | | |
| **Variable** | **Random Forest** | **MARS** | **Elastic Net** | **Linear Regression** | **Average** |
| Production Day | 1 | 1 | 1 | 1 | 1.0 |
| Glucose Trigger Limit | 4 | 2 | 3 | 5 | 3.5 |
| Temperature | 5 | 3 | 5 | 4 | 4.3 |
| Media Type (=B) | 8 | 6 | 2 | 3 | 4.8 |
| Target Cell Seeding Density | 3 | 5 | 7 | 6 | 5.3 |
| Feed Added (%) | 2 | 4 | 8 | 7 | 5.3 |
| Feed Type (=3) | 9 | 7 | 4 | 10 | 7.5 |
| Vessel Volume | 11 | 7 | 11 | 2 | 7.8 |
| pH Setpoint | 7 | 7 | 9 | 8 | 7.8 |
| Dissolved Oxygen (DO) | 6 | 7 | 10 | 9 | 8.0 |
| Strain (=S) | 10 | 7 | 6 | 10 | 8.3 |
| Feed Type (=2) | 11 | 7 | 11 | 10 | 9.8 |

Furthermore, we perform a comprehensive feature ranking based on the variable importance results as the output of employed ML models. Given that GBRT showed inferior performance compared to other models, we exclude its variable importance rankings from consideration. Since each model generates importance values on different scales, we employ a simple rank-average method to standardize the comparison. The aggregated rankings are shown in Table 7, where a rank of 1 is assigned to the most significant variable and a rank of 12 to the least significant, out of a total of 12 variables. Variables assigned the same rank are considered to have comparable importance within the model's variable importance hierarchy. The findings demonstrate that the five most critical factors in determining titer production are production day, glucose trigger limit, temperature, media type B, and target cell seeding density. Furthermore, our results identifying temperature as a crucial factor in predicting protein titer are consistent with the findings reported in previous studies [11].

# Conclusion

This research successfully identified the most accurate model for predicting protein titer, utilizing insights from a variety of models to enhance titer production strategies. The methodology encompassed imputing missing data, training diverse model types, analyzing variable importance metrics, and conducting ANOVA to understand the relationship between key variables and titer production. The proposed approach has not only pinpointed effective strategies for increasing titer yield but also paved the way for applying these insights across similar biomanufacturing contexts. Future research could focus on refining predictive models by incorporating other real-world datasets, exploring additional machine learning algorithms, and further investigating the interactions between significant variables. Collaborating with industry partners for practical validation and exploring the scalability of proposed solutions would also be valuable. Additionally, integrating advanced analytics, such as deep learning, could uncover novel insights into bioprocessing dynamics, leading to breakthroughs in biomanufacturing efficiencies.

Acknowledgements

The authors would like to thank Cytovance Biologics for their generous contribution of biomanufacturing datasets and their close collaboration during the research project. The authors are further grateful for the support by “OKC ED Fd: OU Biotech Core Facility”, Oklahoma City Chamber of Commerce, U.S. Economic Development Administration (Award #: 08-79-05677).

References and Citations

[1] A. Aijaz *et al.*, “Biomanufacturing for clinically advanced cell therapies,” *Nature Biomedical Engineering*, vol. 2, no. 6, pp. 362–376, Jun. 2018, doi: 10.1038/s41551-018-0246-6.

[2] S. Nosaki and K. Miura, “Transient expression of recombinant proteins in plants,” in *Methods in Enzymology*, vol. 660, Elsevier, 2021, pp. 193–203. doi: 10.1016/bs.mie.2021.04.021.

[3] N. K. Tripathi and A. Shrivastava, “Recent Developments in Bioprocessing of Recombinant Proteins: Expression Hosts and Process Development,” *Frontiers in Bioengineering and Biotechnology*, vol. 7, p. 420, Dec. 2019, doi: 10.3389/fbioe.2019.00420.

[4] K. H. Khan, “Gene Expression in Mammalian Cells and its Applications,” *Advanced Pharmaceutical Bulletin; eISSN 2251-7308*, 2013, doi: 10.5681/APB.2013.042.

[5] R. O’Flaherty *et al.*, “Mammalian cell culture for production of recombinant proteins: A review of the critical steps in their biomanufacturing,” *Biotechnology Advances*, vol. 43, p. 107552, Nov. 2020, doi: 10.1016/j.biotechadv.2020.107552.

[6] A. Matte, “High-throughput, parallelized and automated protein purification for therapeutic antibody development,” in *Approaches to the Purification, Analysis and Characterization of Antibody-Based Therapeutics*, Elsevier, 2020, pp. 181–198. doi: 10.1016/B978-0-08-103019-6.00009-6.

[7] K. A. R. Packiam *et al.*, “PERISCOPE-Opt: Machine learning-based prediction of optimal fermentation conditions and yields of recombinant periplasmic protein expressed in Escherichia coli,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 2909–2920, 2022, doi: 10.1016/j.csbj.2022.06.006.

[8] D. Bonanni *et al.*, “A Deep Learning Approach to Optimize Recombinant Protein Production in Escherichia coli Fermentations,” *Fermentation*, vol. 9, no. 6, p. 503, May 2023, doi: 10.3390/fermentation9060503.

[9] B. Wang, W. Xie, T. Martagan, A. Akcay, and B. van Ravenstein, “Optimizing Biomanufacturing Harvesting Decisions under Limited Historical Data,” 2021, doi: 10.48550/ARXIV.2101.03735.

[10] S. V. Buuren and K. Groothuis-Oudshoorn, “**mice** : Multivariate Imputation by Chained Equations in *R*,” *Journal of Statistical Software*, vol. 45, no. 3, 2011, doi: 10.18637/jss.v045.i03.

[11] J. Xu *et al.*, “Systematic development of temperature shift strategies for Chinese hamster ovary cells based on short duration cultures and kinetic modeling,” *mAbs*, vol. 11, no. 1, pp. 191–204, Jan. 2019, doi: 10.1080/19420862.2018.1525262.