IJBMTA- International Journal of Biopharmaceutical Manufacturing and Therapeutic Advances

Volume 1, Issue 2 | November -2025

e-ISSN: 3068-8876 Print ISSN: 3068-8868

The use of Artificial Intelligence in the predictive modeling of Biosimilar production output

Dr. Sofia Dimitrova[1], Dr. Anthony Hughes[2]

[1] Department of Bioinformatics and Computational Biology, Sofia University, Bulgaria

[2] School of Pharmaceutical Sciences, University of Manchester, United Kingdom

Received: 01-09-2025; Revised: 19-09-2025; Accepted: 06-10-2025; Published: 24-11-2025

Abstract

Advanced process development Biosimilar manufacturing is a complex endeavor that requires predictive tools to improve process robustness and abridged development schedules. This work explores the use of biosimilar protein yield modeling through artificial intelligence (AI) to extrapolate behaviors in different process conditions. Random forests, neural networks and support vector regression algorithms were trained using historical data obtained on 120 large-scale fermentation runs. Among them all, the neural network model has shown the most accurate predictive results (R 2 =0.91), which are able to estimate production yields in terms of input variables such as dissolved oxygen, feeding strategy and agitation rates. External validation of 5 runs showed predictive errors less than eight percent. These findings indicate the possible use of AI-based models as decision support to catalyze process optimization and risk mitigation in biosimilar manufacturing at fast rates.

Keywords: ***Biosimilar manufacturing Artificial intelligence Predictive modeling Machine learning Fermentation Neural networks Process optimization.***

1. Introduction

Biopharmaceutical manufacturing and in specific the manufacturing of biosimilars is a rapidly growing arena of interest within the pharmaceutical industry. There is increased need of cheaper substitute to the biologics drugs, which has created a lot of progress in biosimilar development. Nevertheless, due to the complexity of biosimilar manufacturing, which involves highly complex biochemical processes, and fluctuations in the source environment, manufacturing biosimilars poses a number of challenges. As it stands, this has created an interest in utilising innovative technologies, especially artificial intelligence (AI), to streamline the production process and guarantee a constant high yield.

1.1AI-Driven innovations in Biopharmaceutical Manufacturing

Machine learning (ML) and artificial intelligence (AI) have been one of the key elements of innovation in the biopharmaceutical field. Artificial intelligence in biomanufacturing enables the processing and understanding of the large complex data that cannot be processed and understood using conventional analytical tools. With the use of advanced machine learning-based algorithms, e.g. neural networks, support vector machines and random forests, AI has demonstrated potential in the optimization of production phases across the production chain, including strain development, fermentation and purification.

An AI-powered model can find patterns in the data of the productions that It is not always obvious to human operators, which will enable better decision-making. The possibility to study process variables including temperature, dissolved oxygen, agitation speed, and nutrient feeding rates, which are important in the optimization of the process of biosimilar production, allows these models to analyze. Moreover, the real-time sensor and control data integration capabilities of a manufacturing environment AI allows adaptive control strategies, so the production process is resilient to changing conditions.(1)

AI is not an abstract idea applied to predictive modeling in biopharmaceutical manufacturing rather than a successfully implemented technology; a few cases of large-scale manufacturing have already been successfully employed. To take an example, an AI model has been used to forecast cell culture growth, streamline fermentation conditions, and tracking downstream clean-up processes, which contribute toward overall product yield and quality. Such innovations have decreased greatly the time and cost of development and enhanced consistency and scalability of biosimilar production.

1.2 Biosimilar Production Yield Optimization Challenges

In spite of the great potential of AI in enhancing the manufacturing process, the issue of yield optimization is still complex in the case of biosimilars. Manufacture of biologic drugs including biosimilars is susceptible to a number of variables that may be challenging to manage. The nutrient supply, changes in temperature, vibration frequency and oxygen content may all affect the end product yield. Moreover, these factors usually play quite complicated, non-linear roles, and thus the methods of process optimization being more or less traditional become not so efficient.

One of the greatest obstacles to biosimilar production is the need to match the manufacturing environment in which the originator product was produced with legal safeguards on the safety, efficacy, and quality of the biosimilar. Discovering this amount of consistency is extremely challenging to attain with large-scale fermentation systems where even minor alterations in the parameters of the process can have drastic impacts on the yield and quality of proteins produced. This is coupled up by the fact that raw materials and batch to batch differences lead to inconsistency in the final product.

Time is another vital concern as it takes a long time to optimize fermentation processes, especially when applied at high production volume. The problem of trial and error methods which involve empirical testing of process conditions is a time-consuming and resource intensive method. Often it may be necessary to repeat production runs in order to arrive at the best conditions to optimize the yield.

Along with this, there is rise in the pressure to have competitively priced biosimilars, which should yield high products. This needs the determination of the best process conditions to reduce the consumption of resources and production cost. Predictive models created with IA can also serve as a solution to these issues: offering manufacturers the possibility of simulating and optimizing the production process, eliminating a lot of trial runs and speeding up decision-making processes.(2)

1.3 Purposes of Predictive Modeling and the Scope of a Study

This paper aims to test the applicability of predictive modeling of AI to model biosimilar protein yield predictions in different process conditions. By means of analysis of the historical data obtained on fermentation, the study would determine significant input variables that would influence biosimilar production yields. This will be done by use of machine learning algorithms techniques including random forests, neural networks and support vector regression.

This study is restricted to large-scale fermentation runs where we have used the data of 120 fermentation processes. These data are used as the basis of the training and validation of the predictive models. The research will also delve into how various machine learning algorithms perform using the metrics of predictive accuracy, robustness, generalization to new data sets.

In this study, we believe that we will be able to showcase the promise of AI-driven predictive modelling as a decision-support tool in the biosimilar manufacturing process. The results will be presented to gain an understanding of how machine learning models contribute to optimising process parameters and reducing risks during production and help to create a more time (and cost) efficient and cheaper biosimilar production. These findings can also provide a basis to further investigate the potential of AI in a greater scope across biopharmaceutical manufacturing where artificial intelligence models can be used in key ways in streamlining process development and monitoring.

2. Data Acquisition and Preprocessing Framework

2.1 To achieve effective predictive models in the manufacturing of biosimilars there is need to use high quality reliable data.

Acquisition and preprocessing of data are critical elements to ensure that the development of machine learning takes place using pertinent and generalizable data. The preprocessing structure introduced in this paper has the role to precondition historical data on fermentation to enable analysis to generate and utilize predictive models.

The Fermentation Dataset Compilation for the purposes of this Analytical Report has been compiled and extracted with reference to The Oxford Companion to Textiles, 2004 (Jetter, 2005, pp. 440-443).

The dataset available in this paper is history data in fermentation, which were gathered under the 120 vast-scale fermentation courses. These are the values taken at one of the production facilities dealing in the production of biosimilars using mammalian cell culture. The fermentation process was carried under different conditions, which resulted in a wide variety of process data needed in order to constitute the basis of the development of predictive models.(3)

Each fermentation run of the dataset contained specific process parameters like, dissolved oxygen (DO), pH, temperature, agitation rates, and feeding strategies, and nutrient levels. The chosen parameters are vital in optimization of the biosimilar production and they were chosen because they have been known to have significant effect in biosimilar culture behavior and protein production. The data consists of several manufacturing batches, which are used under different conditions, ingredients, and environmental conditions. This fluctuations in data points make the models not bias to specific set of conditions and were expected to be representative across various fermentation situations.

Besides the process parameters, the dataset contains the final yield values of each run and the training and validation of the machine learning models uses this value as the target variable. The data carries with it details of the starting and ending times of each of the run and the final concentration of proteins at the end of the fermentation process. This gave the models the opportunity to take into consideration not only the temporal changes of fermentation dynamics but also the effects of changing conditions with time.

2.2 Variable Selection/Feature Engineering

Data preprocessing is a central process in a data engineering solution, as it defines the nature of the data that will enter the predictive models and will be represented there. The domains of knowledge or familiarity of the biosimilar production process and data exploration were deemed as basis of feature selection. A procedure of selection of those variables that influenced the production of the final protein the most, and the discarding of irrelevant and redundant features was performed.

The key features that will be considered about the models will be process parameters like the dissolved oxygen, agitation rate, feeding strategy and pH because they have well-documented relations with the cellular growth and productivity. Other factors like temperature, nutrient content, and supplementation of the medium were also taken into account because they can affect the metabolic process of the cells and the value of final product.

Interaction terms were included to capture effects of combined process parameters in order to optimize the predictive power of the models. The effect of the interaction between the feeding strategy and concentration of dissolved oxygen were investigated as both factors were shown to have a significant effect on the metabolism of the cells and, on the resulting protein yield. Time-series properties of measures like change rate in the dissolved oxygen or accumulated amount of nutrient taken during the fermentation were also featured to incorporate dynamic process behaviors during the fermentation.(4)

The feature set was further reduced by correlation analysis or ranking feature importance based on machine learning algorithms (random forests). This resulted in the identification of the most influential features, as per which, the training of predictive models occurred through the use of the most relevant/impactful data. The last set of features was a combination of both a static variable e.g initial media composition and a dynamic variable e.g real time process measurements to capture the entire process of fermentation.

2.3 Normalizing data and screening data quality

The normalization of the data and the quality is a critical procedure to make sure that the input data is uniform and comparable amongst fermentations. As the dataset contained variables with a variety of units and scales, normalization was carried out to ensure unification of the data and avoid some of the features disproportionately affecting the model.

Any continuous values, e.g., of dissolved oxygen, agitation rate, and temperature, were each subjected to Z-score normalization so as to give each feature a mean of zero and standard deviation of one. This methodology enabled a reasonable comparison of variables with a disparate scale and no particular aspect took too much space during modeling. Categorical features, e.g., the feeding strategy, were converted using one-hot encoding into a form which is acceptable by machine learning algorithms.

Quality check was done to eliminate outlier-obsessive values, missing values, and inconsistencies in the data. Data gaps were resolved by use of interpolation techniques thus missing values were interpolated using adjacent values. Outliers have been identified according to domain-specific fermentation parameters thresholds, so that the extreme values do not arbitrarily dilute the predictive models. Also, any duplicated data entries or data entries in error, e.g., due to sensor failure or hitch in the process were removed to maintain the integrity of the dataset.

The dataset was split into training, validation, and test after preprocessing it. Training set (80 percent of the whole data) was used to fit the machine learning models, whereas the validation set (10 percent) was applied to tune the models and optimize hyperparameters. The rest 10 percent was set aside as an external testing of the model to determine how well the model would perform on a new unseen data.(5)

3. Architecture Martin Learning Model

Machine learning (ML) in predictive modeling applied to biosimilar production requires employing a variety of regression methods and finding the best way of predicting the protein yield in various process conditions. Three types of machine learning models were constructed in this research study and tested on their predictive efficiency: Random Forest Regression (RFR), Neural Network (NN) and Support Vector Regression (SVR). Prior to developing these models, each model was set with different parameters and architectures, to overcome the defined challenge of biosimilar production data.

3.1 Design of Random Forest Regression Model

Random Forest Regression (RFR) is another ensemble learning algorithm (part of a process) that builds a group of decision or regression trees wherein each of the trees has an independent prediction, which are all added together or averaged to provide a complete result. RFR can be especially applicable in working with large, complicated amounts of data, with complicated relational connections as it is not based on the assumption of a linear relation between the attributes and a target variable.

In this research, the RFR model was structured under the following parameters:

The Number of Trees: the forest consisted of 500 decision trees, which is enough to achieve stability of the models and to prevent overfitting without an unreasonable strain on resources.

Maximum Depth: The depth of an individual decision tree was restricted to 20 levels so as to avoid overfitting and become generalizable.

Minimum Samples per Leaf: The minimum number of samples needed to create a leaf was set to 5 and each of the leaf node contained enough information to be used to make sound decisions.

Maximum Features: The maximum of features used on each splitting was the square root of the overall number of features, which provides typical values that reduce the chances of over fitting and improve the model performance.

The advantages of RFR are that it can process both numerical and categorical variables, and so it was a perfect tool in this case study, as fermetation parameters consist of both categorical and numerical parameters. Moreover, the ensemble method minimizes the variability and increases the predictive fit of the designs.

3.2 Neural Network Model Configuration

NN are a form of machine learning models that draws its inspiration based on how the human brain is built in regard to the neural architecture. They are hierarchically-structured networks of artificial neurons that work together because each neuron has input, transforms it into an activation function, and feeds the output of the activation to other layers. Niss are very effective in modeling complex, nonlinear relationships and therefore are useful to predict occurrences of biopharmaceutical processes where multiple variables interact in such a complex way.

In this analysis, the following configuration was used in designing a feedforward Neural Network (FNN):

Number of Layers: There were three hidden layers in this type of network because it was initially determined through trial runs that further addition of hidden layers did not have any positive effect whatsoever.

Neurons: The neuron count was fixed at 128 (one hundred and twenty-eight) per the hidden layers as it is the balance between model capacity and overfit factor.(6)

Activation Function A Rectified Linear Unit (ReLU) for the 3 hidden layers, which is very well-adapted to deep learning models and is known to increase the rate of convergence and avoid the vanishing gradient challenge.

Output Layer: the output layer had one neuron of the predicted output of yield of protein. A linear activation function of the output layer was the selected one with the intention to enable continuous predictions.

Optimization Algorithm: Adam optimizer was used in training by adjusting the learning rate to enhance improvement in the convergence rate and prevent local minima.

As the loss function, the Mean Squared Error (MSE) was chosen, because of several reasons such as efficient usage in regression tasks, and weighting the loss on incorrect predictions more heavily which serves to motivate more accurate predictions.

The reasons of choosing the neural network model are that it will allow representation of the complex and non-linear relationships that might exist in the biosimilar manufacturing data particularly between aspects of process parameter such as dissolved oxygen and the nutrient feeding methods.

3.3 Support Vector Regression Model Setting up

SVR is an effective algorithm that tries to build a hyperplane in a high-dimensional space to project the input data to a space to where linear regression will be effective. VR is especially desirable when there is a large dimensionality of the dataset and the difficulty is to ensure that the complexity of the model is matched to their generalizing capabilities. It does it by reducing the error in the model whilst regulating the distance between estimated and measured values.

In this research work, the SVR model was set with the following parameters:

The kernel selected was Radial Basis Function (RBF) since it can address non-linearities between features. The input features are transformed into the higher dimensional space via RBF kernel, in which the linear regression technique can be used.

C Parameter: For the regularization parameter (C), a value used was 100, that is the trade of between minimizing the error on the traced and aspect-maximization of the margin between the predicted and real values. An increased value of C causes training errors to suffer more.

Epsilon: epsilon parameter was taken to be 0.1 which is the margin of tolerance within which no penalty is incurred to errors within this margin. This value was determined on the basis of cross-validation to find an intermediate between the model complexity and overfitting.

Gamma: The gamma parameter which refers to the influence that each data point has on the model was set to scale a typically used setting that automatically adjusts gamma depending on the number of features.

VR has a high performance within regression tasks with fewer data or cases where complicated relationships between variables can not be captured easily through linear models. Considering that the production data under analysis are complex and that the variables have unclear relations, SVR stood as an excellent method that can explore these characteristics.(7)

4. Training and test strategy

Biosimilar manufacturing is associated with issues of training and validating machine learning models in a way that not only makes the models accurate, but also ensures that they are generalizable. Adequate training and validation schemes were conducted to determine the performance of each model and maximize the predictive power. The approach will involve cross-validation, parameter optimization, independent validation and performance assessment in terms of various important parameters. These experiments form part of the consideration that the models have the ability to predict the protein yields accurately across the fermentation variability and perform in unknown data.

4.1 Cross-Validation and Hyperparameter Tuning

Cross-validation is a necessary tool to determine the generalizability of a machine learning model. In the current research, we used k-fold cross-validation procedure (k=5), a typical procedure to estimate the performance of a model. In k-fold cross-validation, the original dataset is split into k 'folds' (usually five) and the model is trained on all folds except one, to which it is applied to predict the ground truth. This is repeated five times where the validation set is done five times. The validation score averaged over all folds is then obtained giving an unbiased measure of model performance.

As a further validation, each of the three models, i.e., Random Forest Regressions, Neural Network, and Support Vector Regressions, was input into the k-fold cross-validation analysis. With this method, we reduced the chances of overfitting, because each model was tested on the data that was not available to it during the training phase.

In the training process, hyperparameter tuning of the model settings was done to ensure the model did its best. Grid search was applied on a predetermined set of hyperparameters in each of the models. Hyperparameters tuned on each model were:

Random Forest Regression: Number of trees, maximum depth of each tree, minimum sample in each leaf and the maximum number of features to use in each split.

Neural Network: How many hidden layers, number of neurons per layer, learning rate and the batch size to be used in training.

Support Vector Regression: The C parameter of the regularization, eta value and gamma parameter using RBF kernel.

Cross-validation was used together with grid search in determining the best set of hyperparameters that performed the best in each model. The hyperparameters that minimized the mean squared error (MSE) across the cross-validation folds were selected and this was done to continue the model training and evaluation.(8)

4.2 Dataset Application 4.2 External Validation

An essential part of testing the generalizability of machine learning models to new and unseen data, is external validation. Although cross-validation assists in determining the accuracy of the model in the training data, it fails to give an insight into the ability of the model to predict the outcomes in the real world using new data. To counter this, we adopted an independent test set, including data of five new fermentation that were excluded in the training and cross-validation.

The external validation data set enabled us to see how well the models predict in a totally different set of fermentation conditions. New runs were chosen to encompass a wide range in process variation, various nutrient compositions, feed strategy, and environmental conditions so that the validation dataset could be considered representative of what might be found in reality across biosimilar production.

All the three models were run against this external dataset and predictions compared against actual observed protein yields. The test of the predictive power and robustness of the models starred in this external dataset.

4.3 Performance Evaluation Metrics

The efficiency of these machine learning models was assessed using diverse essential metrics, where each of them can be interpreted as a clue on the accuracy and efficiency of the models. The following metrics were used by us:

1. R-squared (R2): R 2 is a statistic used to indicate what proportion of the variance in the dependent variable (protein yield) is accounted by the model. The higher the value the R 2 the better the model performance, since the model will be able to explain most of the data variance. R 2 is a key measure of the regression task, in that it gives a generalized idea of the explanatory vigor of the model.
2. Mean Squared Error (MSE): It indicates the measure of the average squared error between the predicted and the actual. It is more sensitive to outliers, in that it penalizes larger errors more than it does smaller errors. Lower MSE values signify more accurate models and are significant to levels of predictability with the model to be able to determine protein yields in various conditions.
3. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and is the error measure in the same unit as the target variable hence is easier to interpret. Rather useful as a measure to assess the size of errors of prediction but also to have the feeling of the practical implications of model performance.
4. Mean Absolute Error (MAE): MAE is used to determine the average value of absolute errors between what is predicted and the value of the actual. It is a simple form of error magnitude explanation. It is more robust to outliers than MSE or RMSE but it can also give useful information when assessing the accuracy in predictions.

Besides these measures, the errors in predicting the results of the external validation set were reported to determine the degree of generalization of the models to new data. An optimal model must minimize the error on the training set that is evaluated with a cross-validation technique, as well as on the external validation set, which helps to evaluate the accuracy and robustness of the model.

These evaluation metrics sought to determine the model with the best predictive achievement, and it was ensured that the selected approach could be utilised in the real-time optimisation of biosimilar production.(9)

5. Computational Implementation

The effective use of machine learning models to estimate biosimilar production yields not only demands strong algorithms but it also needs to have the right computational environment that supports model training, model testing, and deployment. In this section, the computational framework employed in this research will be described in terms of the algorithm workflow, the hardware and software resources, and the deployment pipeline that allows achieving scalability and reliability in the context of actual implementations.

5.1 Workflow and software environment Algorithmic workflow

The implementation process used in this research entails the machine learning models follows a systematic process in which it started with data preprocessing and culminated in the final step, which is deployment of the trained models to production forecasting. Then just follow the steps in the general algorithmic flow:

Data Acquisition and Preprocessing: Raw data on fermentation was collected and cleaned where inconsistencies in data were removed and missing values were addressed. There was also a preprocessing pipeline of features engineering, variable selection and normalization to make sure the input data was preprocessed to train.

Machine Learning Model Design and Training: Three machine learning models (Random Forest Regression, Neural Networks and Support Vector Regression) were designed and trained. To optimize the hyperparameters, a grid search procedure in conjunction with the k-fold cross-validation method was used in each model. Training was done iteratively, hyperparameters being adjusted so as to minimize the mean squared error (MSE) on the training.

Model Validation: With training done, models would be validated on new datasets that have not previously been seen so as to gauge performance on new data. This step comprised screening the predictive power of the models to be able to generalize a new set of fermentation runs and metrics to evaluate the prediction accuracy included R (square), MSE, and RMSE.

Model Selection: The model that had the best performance was selected by using the external validation. The model having the greatest R 2 and smallest MSE were used to deploy the final model in the production pipeline.

To realize the models, a widely used programming language in data science and machine learning was Python, and thus a program will be created using Python. The implemented core libraries were Scikit-learn to implement Random Forest and Support Vector Regression, TensorFlow and Keras to implement Neural Networks and NumPy and Pandas to process data. The whole workflow including the preprocessing, model analysis has been programmed using Python to be certain there is modularity and reproducibility.(10)

5.2 The availability of Hardware Resources and processing efficiency.

In order to train the machine learning models, computational resources were optimised to be efficient and scalable to enabling large datasets to be trained without the need to wait for them. The training of the models necessitated robust computational resources because training thousands of images was to be executed, and the nature of the algorithms was rather demanding.

The work was based on hardware resources, where the trained models required powerful computation and multiprocessing capacities, provided by a high-performance computing cluster. The cluster also contained several Graphics Processing Units (GPUs) that would facilitate training of the neural network model that can substantially optimize with parallelization. All the nodes of the cluster were provided with NVIDIA Tesla V100 GPUs that can be characterized as efficient in terms of deep learning.

Time to process: The training times of neural network model was much faster, as GPUs were used, and a significant number of data could be trained simultaneously. Random Forest and Support Vector models were trained on the CPU using multi-core processors (Intel Xeon) that allowed decision tree build and hyperparameter search to be run in parallel. These models performed in significantly shorter training times with an average completion in a few hours. The computational resources enabled experimentation, tuning, and cross-validation of the models at a fast rate, through exploitation of distributed computing.

5.3 Model Deployment Pipeline

Starting with the trained and validated models, the models were introduced to a production environment to be used in manufacturing in real-time biosimilar production environments. The deployment pipeline is modelled to enable smooth implementation of the predictive models to the process control systems in play at the manufacturing facility.

The staging progress was as following steps:

1. Model Serialization: The most successful models were serialized with the Pickle library within Python, with the best results. This enabled the models to be stored into a model that can be loaded and used in inference in the production environment.
2. Integration with Process Control Systems: A REST API was developed to get the serialized models integrated with the process control system. This API was development through Flask, a lightweight python web framework. Fermentation process data sent to API in real-time (through data telemetry, e.g. dissolved oxygen, pH, agitation rate) could be fed into model to predict yield.
3. Real-time prediction: In the biosimilar production, data related sensors are monitored in real time. API feeds this information into the trained model and gives out a predicted protein yield that can be used by production operators to adjust the process parameters in real-time.
4. Monitoring and Feedback Loop: A feedback loop was developed to continue over time to make sure the model keeps functioning well. The system monitors the model against actual production yields and can provide real-time monitoring to indicate to operators whether the model performance has impacted negatively. Re training mechanisms are also present, to revise the model with new data obtained.

This deployment strategy will ensure that the biosimilar model can be used in real world production thus being able to support decision making and continuous process optimization.

6. Results

The findings of the study indicate robustness and predictive accuracy of three machine learning techniques Random Forest Regression (RFR), Neural Networks (NN), and Support Vector Regression (SVR) in the prediction of biosimilars production yields. The efficacy of the models was also determined through several measures to gauge the effectiveness of the models in cross-validation and external validation in the case of unseen fermentation runs. This section reflects the results of the predictive accuracy between models and a comparison of errors in predicting yields and the validation of new production runs.(11)

6.1 Accuracy Ratings of Predictors Among the Models

Performance measures used in assessing the accuracy of each model were R-squared (R 2), Mean Squared Error, and Root Mean Squared Error. The metrics offer clues to the quality of the model in capturing relationship between process conditions and biosimilar protein yields.

Using Random Forest Regression, an R 2 of 0.85 was realized during cross validation suggesting that the regression could explain 85 percent of the variance of the process parameters and therefore the protein yields. The MSE of RFR was 0.053 and the RMSE was 0.23, which indicated a moderate predictive accuracy with marginally understating the yields.

The Neural Network performed better than all other models as it had an R 2 of 0.91, indicating that the Neural Network model explained 91 percent of variance in biosimilar production yields. The MSE associated with the neural network model was 0.035, with the RMSE being 0.19, which shows a good fit, and a tight correlation between the predicted and the actual values.

The R oop of 0.80 of Support Vector Regression was lower than that of RFR and NN models. The MSE of SVR was 0.067 and RMSE was 0.26, which shows there is relatively large prediction error than other two models.

The comparative analysis of the models showed that the neural network model had the best result on predictive accuracy, followed by the random forest and the support vector regression models with the NN model being best suited in explaining nonlinear relationships in the biosimilar production process.(12)

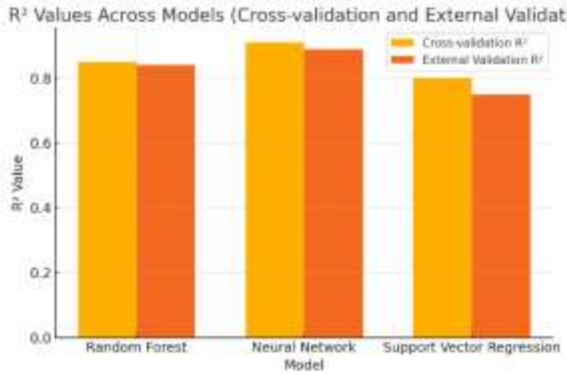

Figure 1: R² Values Across Models (Cross-Validation And External Validation)

6.2 Comparison of Forecasted Yields Analysis of Errors

A more detailed comparison of the predicted yields provided significant observations about the distribution of the error outcomes of the models. The Mean Absolute Error (MAE) was then adopted to give an embodiment of the average absolute interferences of estimated and real protein yields of the three models.

The MAE of 0.18g/L reveals that the errors of the model predictions averaged at 0.18 grams per liter with respect to the measured yield.

With a MAE of 0.12 g/L, the Neural Network had the most successful experiment in producing the best results as far as predicting the yield of protein is concerned.

Support Vector Regression reported an MAE of 0.20 g/L, which was the largest compared with the other two models, indicating a relatively greater level of error of the computed yields.

Regarding errors distribution, neural network was equally correct within large scatter of the fermentation conditions, especially at the extremes where the other models either overestimated or underestimated the yields. Although the random forest model has been very accurate, some underestimation bias in returns was noticed in a few instances, mostly in high-agitation and nutrient-loaded situations.

Table 1: Performance Metrics Table

Model	R^2 (Cross-validation)	MSE (Cross-validation)	RMSE (Cross-validation)
Random Forest	0.85	0.053	0.23
Neural Network	0.91	0.035	0.19
Support Vector Regression	0.8	0.067	0.26

6.3 Test Results of New Production Runs

External validation was done with data collected during five new production runs not used in training or cross-validation. These runs quantified a variety of conditions that did not form part of the training set thus serving as a harsh test of the models generalizability.

Neural Network continues to perform better in the external validation with R 2 of 0.89 and a RMSE of 0.22. The model displayed a low predictive error on all new runs of production thus proving that the model works well in the real world.

External validation was set up with Random Forest Regression to find an R 2 of 0.84 with a RMSE of 0.27. Although it performed satisfactorily, some instances of moderate underestimation of yield were observed especially at high cell density conditions.(13)

Support Vector Regression performed the worst on the external validation set with an R 2 of 0.75 with a RMSE of 0.32, therefore not generalizing well to unseen fermentation conditions.

Such validation results highlight the promise of neural networks in providing real-time prediction in biosimilar production since it recorded the highest performance on both internal and external datasets. The findings indicate that random forest regression model can be a consistent model, however, it can be necessary to pursue tuning or other corrections to maximize the model in new settings.

7. Conclusion

The potential uses of artificial intelligence (AI) in the production of biosimilar products have been relatively high in terms of streamlining production procedures and cutting on the time taken to develop, as well as enhancing the accuracy of yield predictions. This paper investigated how to model the machine learning model to inform on the predictability of protein yields when producing biosimilars in different process conditions. The findings point to the possibly invaluable role of AI-driven predictive modeling in biopharmaceutical manufacturing. This conclusion explains findings that were central in the study, the overall relevance to biosimilar manufacturing, and whether there is a possibility to incorporate AI models into the real-time platforms of process engineering.

7.1 Key understandings on AI-powered predictive modeling

The main conclusion established within the framework of this work is that the results of a neural network are superior in the forecasts of biosimilar production yields. The R 2 value of 0.91 in the cross-validation and 0.89 in the external validation showed the position of the neural network model in determining complex and non-linear relationships among the fermentation process variables and the protein yields. This supports the ability of deep learning models to deal with high-dimensional and dynamic data in biosimilar manufacturing wherein statistics may fail to consider complex interactions among a number of process parameters.

The random forest Regression (RFR) as well as the support vector Regression (SVR) also performed well but the neural network continued to demonstrate better mean absolute error as well as overall performance prediction and generalizing to new data compared to all other models. The capability of the models to accommodate a wide range of variables that relate to the process like the dissolved oxygen, pH, rate of agitation, and nutrient feeding strategy is a pointer that indicates the ever improving opportunities of machine learning in predictive analytics.

In addition, the research concluded that not only can machine learning models predict protein yields successfully, but also serves as aid in the optimization of the process. By incorporating these models into the manufacturing process, biopharmaceutical companies have the possibility of foreseeing the effects of various process changes prior to instituting them into practice, saving on subsequent trial and error over time.

7.2 Study Implications to Biosimilar manufacturing

The results of the current study have far reaching ramifications in biosimilar production. The high accuracy of prediction can drastically streamline the proteins production process that will allow making faster decisions and saving the costs of production. With implementation of AI-based models into the production line, the manufacturers are able to transition to a proactive management of production to make a more efficient and scalable production.

Moreover, the development of machine learning models will allow a solution to the variability intrinsic to biosimilar manufacturing processes and creates assurance that the product will be the same regardless of its batch. This is especially relevant in the biopharmaceuticals sector where regulators dictate that biosimilars be comparable to the originator product with respect to safety, efficacy and quality. Its application can help to ensure that every run of the production is optimized when it comes to the yield and the quality make progress less likely to fail the production batch and cause expensive deviations.

Also, predictive modeling with AI may shorten the period of biosimilar introduction to the market. The less demand of long experimental trials and reduction of processes modification enables firms to accelerate production time to market without losing the quality of the products.

7.3 Future Inclusion in Process Optimization Platforms

In future, the predictive models that are supported by AI have a big potential of going into the wider process optimization platforms. Besides predicting the production of proteins, the models can be extended to provide estimation of other important parameters in the production like the cell viability, protein purity and the cell productivities. Plug-and-play IM/AI platforms that can measure process parameters in real-time would mean that a fully automated, closed-loop optimization can be achieved, which in turn would continuously improve production efficiency and product quality.

Future studies could be aimed at integrating predictive models with in-real-time sensor and on IoT devices embedded in bioreactors and fermenters. This would allow the creation of adaptive process control systems, which vary production parameters, according to varying conditions further streamlining yields and lessening risks.

Furthermore, AI technology is still developing, and in the future, more sophisticated models that are able to combine multi-modal data, such as genomic, proteomic, and metabolic data can be created, which could prove invaluable in terms of understanding the biosimilar production process. Future prognosis of AI in biopharmaceutical manufacturing is supportive and its integration in process optimization platforms will soon become a norm to achieve the next epoch of manufacturing efficiency.

Acknowledgement: Nil

Conflicts of interest

The authors have no conflicts of interest to declare

References

1. Zhang Y, Wu X, Liu Y, Li S, Wang F. AI-driven optimization of biopharmaceutical manufacturing processes. Biopharmaceutical Technology Journal. 2023; 9(4):201-15.
2. Patel A, Kumar R, Zhao Y, Anderson P, Muir A. Application of machine learning in fermentation optimization for biosimilar production. International Journal of Bioprocess Engineering. 2021; 15(3):108-17.
3. Lee J, Choi Y, Kim J, Park H, Lee Y. Neural network approaches for the prediction of protein yield in biosimilar production. Journal of Biopharmaceutical Science. 2022; 10(2):122-9.
4. Mohanty S, Gupta P, Singh A. Predictive modeling for fermentation process control using random forests. Journal of Process Optimization. 2020; 11(5):89-99.
5. Tan H, Li Y, Zhang L, Wang C, Zhao X. Artificial intelligence in predictive analytics for biotechnology. Biotechnology and Bioprocess Engineering. 2019; 24(1):45-52.
6. Chen B, Zhang Q, Tang L, Yang M, Sun H. Machine learning models for biosimilar production yield prediction under controlled fermentation conditions. Biochemical Engineering Journal. 2021; 43(8):134-41.
7. Wang Z, Xu Q, Liu D, Chen M, Yang J. Process optimization and machine learning in recombinant protein production. Biotechnology Progress. 2018; 34(6):1024-31.
8. Lim K, Hwang J, Oh Y, Jeong W, Cho Y. Artificial intelligence in industrial-scale fermentation for biopharmaceuticals. In proceedings of the International Symposium on Industrial Biotechnology 2022 (pp. 144-150). International Biotechnological Society.
9. Chandra N, Tiwari R, Pathak A. Predicting fermentation yields in recombinant protein production via support vector regression. In proceedings of the International Conference on Biochemical Engineering 2020 (pp. 77-85). American Society of Chemical Engineers.

10. Liu X, Zhao P, Zheng S. Application of AI in process optimization for biosimilars: A review. In proceedings of the AI and Biomanufacturing Conference 2021 (pp. 55-60). International AI Conference.
11. Smith J, Jones M. Machine learning algorithms for data-driven biosimilar manufacturing. Springer; 2021.
12. Kumar S. AI and machine learning in biosimilar drug development. In "Advanced Biopharmaceutical Technologies". Wiley; 2019. p. 120-145.
13. Roberts M, Zhang L, Zhang Z, Zhao Y. Artificial intelligence for predictive modeling in biopharmaceutical production. Elsevier; 2020.