DL based real-time prediction of product formation in biopharmaceutical manufacturing

Md Nasre Alam,a Sami Ullah Bhat,a Hariprasad Kodamana,a,b,\* Anurag S. Rathorea,b,\*

aDepartment of Chemical Engineering, Indian Institute of Technology Delhi, New Delhi, 110016, India

bYardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, 110016, India

[asrathore@biotechcmz.com](mailto:asrathore@biotechcmz.com), [kodamana@iitd.ac.in](mailto:kodamana@iitd.ac.in)

Abstract

The extraordinary growth in the use of sensors, as well as in our computational abilities, has resulted in a massive amount of data generated in biopharmaceutical manufacturing plants. This, in turn has fuelled researchers to explore the applications of Machine Learning (ML) and Deep Learning (DL) as superior alternatives to the traditional Multivariate Data Analysis (MVDA) methods. For manufacturers of biotherapeutic products, achieving a consistent product titer is paramount. This study explores the application of DL techniques for real-time prediction of growth and product formation in aerobic microbial fermentation in a bioreactor. Process parameters, including dissolved oxygen, airflow rate, revolution per minute, pH, and temperature, were considered input features, and their impact on product formation has been explored, with optical density and protein titer being output variables. Three different DL regression algorithms: U-shaped Network (UNET), Convolutional Neural Network (CNN), and Multilayer Perceptron (MLP), have been incorporated in this study to measure the prediction performance. Their performance is measured in terms of R2, MSE, RMSE, MAE, and MAPE. UNET consistently performed better than CNN and MLP methods. R2 was 0.9094 and 0.8872 for optical density and protein titer prediction, respectively, in the UNET model, followed by 0.8948 and 0.8766 in the CNN model, and 0.8590 and 0.8548 in the MLP model. The study demonstrates how DL techniques can assist in the manufacturing of safe and efficacious biotherapeutic products.

**Keywords**: Deep learning, Multivariate data analysis, Microbial fermentation, Regression

* 1. Introduction

A typical bioprocess consists of multiple unit operations, each intended to perform a specific function. The bioreactor aims to offer a controlled environment to the cells with respect to physical and chemical attributes so that the cells can multiply at the desired rate. However, controlling a fermentation process is non-trivial due to the fact that most biological processes are non-linear in nature (Rathore et al., 2021). Hence, process parameters such as temperature, pH, oxygen transfer, mixing, and substrate concentration need to be controlled to their desired setpoints to ensure optimal performance (Wang & Zhong, 2007).

Over the past decade, modeling of fermentation processes has seen advancements through the application of data-driven models (DDMs), leveraging Artificial Intelligence (AI), particularly machine learning (ML), and deep learning (DL) techniques. For instance, the relationship between inputs and outputs in mammalian cell cultures secreting monoclonal antibodies (mAbs) has been scrutinized using diverse ML methods (Schinn et al., 2021). Several machine learning algorithms based soft sensors have been successfully implemented for correlating input parameters during fermentation with the target variables (Escalante-Sánchez et al., 2018; Wang et al., 2023). These have been applied depending on the online data of different bioprocess parameters. This study tries to further explore the utility and advantage of online data obtained from microbial fermentation process using deep learning frameworks. It is essential to recognize that no single DDM can serve as a universal strategy in output prediction for such datasets. This limitation arises because model performance is intricately tied to factors such as model structure, specific elements of the process, datasets available, and data volume, among other considerations (Von Stosch et al., 2021).

In this study, we introduce a structured framework designed to assess different DL model which utilizes and leverages more deeper understanding of algorithms in capturing intricate pattern of dynamic process data. The proposed framework has been demonstrated to be suitable for real-time monitoring of a microbial fermentation process.

* 1. Materials and methods
     1. Bioreactor cultivation

Recombinant Escherichia coli expressing the protein therapeutics of interest were cultured in Sartorius Biostat Bplus bioreactors (Sartorius, Germany) with a total capacity of 5 liters, employing both batch and fed-batch processes. The initial volume of the medium was set at 1.5 liters. The batch medium was a defined media containing carbon sources, nitrogen sources, and trace metals. Glucose (C6H12O6), glycerol (C3H8O3), magnesium sulphate (MgSO4), thiamine HCl, trace elements, and antibiotics were added separately, with all components dissolved in Milli-Q water obtained from the Milli-Q® water purification system. The pH was regulated before the transfer to the bioreactor for sterilization. Batch medium inoculation was performed with an overnight-grown primary culture of quantity 200 ml. To provide optimal growth conditions, a thermal mass flowmeter was used to control the airflow rate. pH probe (Hamilton®, USA) was used to monitor pH levels. 2 N HCl and 12.5% liquid ammonia were used to maintain pH. D.O. probe (Hamilton®, USA) was used to monitor dissolved oxygen levels. Further, D.O. levels were controlled by adjusting the stirrer speed, air flow rate, and pure oxygen in a cascading manner.

During the cultivation in batch phase, glucose and glycerol concentrations of 10 g/L and 2 g/L, respectively, were kept initially for each bioreactor operation. Upon depletion of glucose, cells transitioned to glycerol as the carbon source after a brief period of adaptation. Exhaustion of the carbon source leads to the dissolved oxygen (D.O.) spike, which signifies the end of batch phase. Glycerol and MgSO4.7H2O concentrations of 400 g/L and 20 g/L, respectively, were added through a peristaltic pump during the fed-batch phase for feeding. Lactose served as an inducer to achieve a working concentration of 5 g/L lactose inside the bioreactor in three pulses.

* + 1. Methodology

Three different models: U-shaped Network (UNET), Convolutional Neural Network (CNN), and Multilayer Perceptron (MLP), were implemented in Python 3.11.4 to model the microbial cell culture data. Data consists of 28 batches each for 5 inputs: (i) dissolved oxygen, DO (%), (ii) air flow rate, F (lph), (iii) revolution per minute, RPM, (iv) temperature, T (°C), (v) pH; and 2 outputs: (i) optical density, OD, (ii) protein titer, PT (g/l). 20 batches were used for training and 8 batches for testing. Training and testing split ratio were 70:30. The statistical study of data is shown in Table 1.

Different model architectures and their workings have been explained below, and implementation details are mentioned in Table 2.

Table 1. Statistical analysis of cell culture data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameters | | #Samples per batch | Minimum value | Maximum value |
| Inputs | DO (%) | 900 | 0 | 112.59 |
| F (lph) | 900 | 18.48 | 143.97 |
| RPM | 900 | 401 | 905.4 |
| T (°C) | 900 | 33.41 | 39.02 |
| pH | 900 | 6.87 | 7.71 |
| Outputs | OD | 1 | 26.6 | 118 |
| PT (g/l) | 1 | 0.09 | 7.8 |

* + - 1. U-shaped Network (UNET)

UNET, characterized by its encoder-decoder structure, is a convolutional neural network (CNN) designed to optimize data efficiency without compromising accuracy and speed. Its architecture features a distinctive contracting and expansive path. The contracted pathway of UNET is designed to identify relevant features within the input image. Spatial sharpness of feature maps is reduced by encoder layer through the convolutional processes. In contrast, the expanding path decodes the encoded data while preserving the spatial sharpness of the image. The decoder layers in the expanding path employ upsampling and convolutional operations, with skip connections aiding in retaining spatial information lost during the contraction, enabling more precise feature localization (Ronneberger et al., 2015). Flattening layer is followed by dense layers are employed after the decoder layers to get the final output.

* + - 1. Convolutional Neural Network (CNN)

CNNs are one of the most widely utilized types of deep neural networks (DNNs), employing layers that facilitate comprehension and interpretation of image or visual data. Serving as a regularized feed-forward neural network, CNN autonomously learns features through the optimization of filters or kernels. A standard CNN comprises three types of layers: input layer, hidden layer, and output layer. The input layer is responsible for supplying model input. The hidden layer gets input from the input layer. Multiple hidden layers may be present depending on the applied model and quantity of data. The number of features increases as the number of neurons increase in each hidden layer. Matrix multiplication over the preceding layer’s output including learnable weights, biases and an activation function of layer contribute to the output of the next layer (Alzubaidi et al., 2021). The output of hidden layers is flattened to feed the dense layer. Finally, the dense layer is sent to the output layer.

* + - 1. Multilayer Perceptron (MLP)

MLP denotes a type of neural network that is extensively employed in ML for classification and regression tasks. Its nomenclature stems from the incorporation of multiple layers of nodes, also known as artificial neurons, interconnected with each other. The framework of an MLP consists of three principal elements: input layer, hidden layers, and output layer. The input layer receives input data and transmits it to the hidden layers, comprising nodes representing features in the incoming data. The number of input data features align with the count of nodes in the input layer. Transformation of input data into an appropriate form for the output layer is performed by hidden layers. Performance for specific tasks can be optimized by adjusting the count of hidden layers and number of nodes in each layer. Hidden layer provides the transformed representation to output layer for getting the output (Popescu et al., 2009).

Table 2. Different models and details of the parameters used

|  |  |  |
| --- | --- | --- |
| Models | Training parameters | Values |
| UNET | Input shape | [900×5×1] |
| Contracting path | #Convolutional layers = 4 |
| #Pooling layers = 4 |
| Bottleneck | #Convolutional layers = 1 |
| Expansive path | #Upsampling layers = 4 |
| #Concatenation = 4 |
| #Convolutional layers = 4 |
| #Learning rate | 0.001 |
| Optimizer | Adam |
| Loss function | MSE |
| CNN | Input shape | [900×5×1] |
| #Convolutional layers | 4 |
| #Pooling layers | 4 |
| Learning rate | 0.001 |
| Optimizer | Adam |
| Loss function | MSE |
| MLP | #Input layer | 1 |
| #Hidden layer | 5 |
| #Output layer | 1 |
| Optimizer | Adam |
| Loss function | MSE |

* 1. Results and discussions

The process parameters were recorded by the Supervisory Control and Data Acquisition (SCADA) system of the bioreactor and stored in an Excel file. The offline data included the target variables: OD and PT for multiple batches of input parameters. Based on the preprocessing of SCADA and offline data, regression results were obtained, and five evaluation metrics: R2, MSE, RMSE, MAE, and MAPE computed for each prediction of outputs (Table 3). The UNET model performed better than CNN and MLP for OD and PT predictions.

Table 3. Different models performance comparison for output variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Output variables | Models | R2 | MSE | RMSE | MAE | MAPE |
| OD | UNET | 0.9094 | 38.8750 | 6.2350 | 6.1250 | 12.0721 |
| CNN | 0.8948 | 45.1250 | 6.7175 | 6.6250 | 13.2219 |
| MLP | 0.8590 | 60.5000 | 7.7782 | 7.7500 | 15.4234 |
| PT (g/l) | UNET | 0.8872 | 0.1906 | 0.4365 | 0.4075 | 63.3211 |
| CNN | 0.8766 | 0.2084 | 0.4565 | 0.4250 | 64.9737 |
| MLP | 0.8548 | 0.2452 | 0.4952 | 0.4113 | 48.8736 |

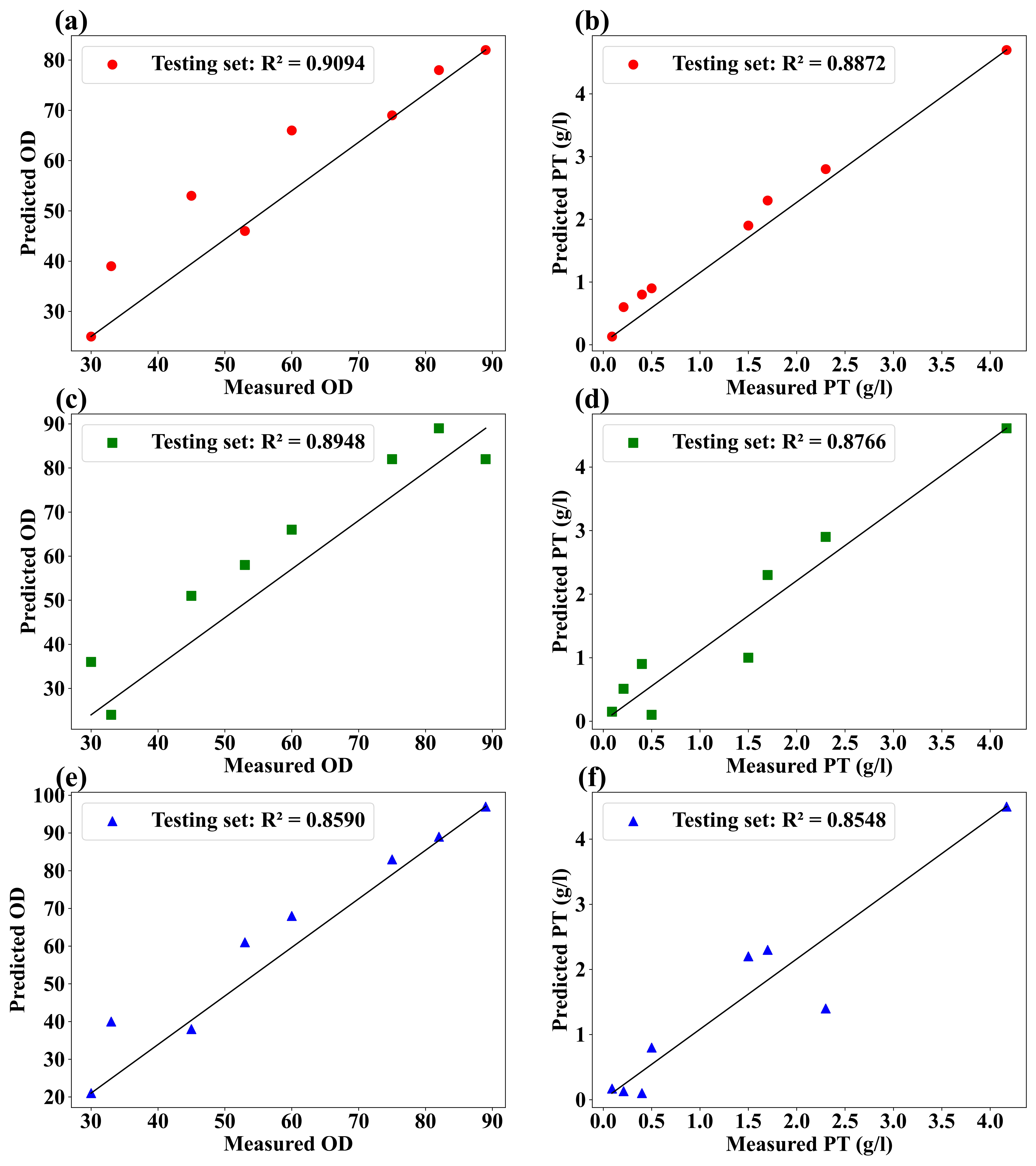


Figure 1. R2 plot for UNET model (a-b), CNN model (c-d), and MLP model (e-f)

The robustness of the implemented model has been demonstrated by the R2 plot for each output variable (Figure 1). R2 plot is shown for 8 batches of OD and PT under testing case for their measured vs predicted values. Higher R2 is consistently obtained for the UNET model for OD and PT predictions followed by the CNN and the MLP models.

* 1. Conclusions

Bioprocessing encompasses vast amounts of data that are collected but seldom analyzed. Process data from all these operations can provide valuable information regarding the process and the product itself. This study explores the application of DL techniques for real-time prediction of product formation in microbial fermentation. Three different models have been applied and compared based on performance. Multiple evaluation metrics were computed to assess model performance, and it was concluded that UNET performed better than CNN and MLP algorithms. In the future study, other ML based algorithms such as gradient boosting model, random forest, etc. can be applied to measure the adaptability and performance. Use of DL models to analyze bioreactor data and predict product formation will facilitate real-time decisions by the operator.

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. <https://doi.org/10.1186/s40537-021-00444-8>

Escalante-Sánchez, A., Barrera-Cortés, J., Poggi-Varaldo, H. M., Ponce-Noyola, T., & Baruch, I. S. (2018). A soft sensor based on online biomass measurements for the glucose estimation and control of fed-batch cultures of Bacillus thuringiensis. *Bioprocess and Biosystems Engineering*, *41*(10), 1471-1484. <https://doi.org/10.1007/s00449-018-1975-3>

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), 579-588.

Rathore, A. S., Mishra, S., Nikita, S., & Priyanka, P. (2021). Bioprocess Control: Current Progress and Future Perspectives. *Life (Basel)*, *11*(6). <https://doi.org/10.3390/life11060557>

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham.

Schinn, S.-M., Morrison, C., Wei, W., Zhang, L., & Lewis, N. E. (2021). A genome-scale metabolic network model and machine learning predict amino acid concentrations in Chinese Hamster Ovary cell cultures. *Biotechnology and Bioengineering*, *118*(5), 2118-2123. <https://doi.org/https://doi.org/10.1002/bit.27714>

Von Stosch, M., Portela, R. M. C., & Varsakelis, C. (2021). A roadmap to AI-driven in silico process development: bioprocessing 4.0 in practice. *Current Opinion in Chemical Engineering*, *33*, 100692. <https://doi.org/https://doi.org/10.1016/j.coche.2021.100692>

Wang, K., Zhao, W., Lin, L., Wang, T., Wei, P., Ledesma-Amaro, R., Zhang, A.-H., & Ji, X.-J. (2023). A robust soft sensor based on artificial neural network for monitoring microbial lipid fermentation processes using Yarrowia lipolytica. *Biotechnology and Bioengineering*, *120*(4), 1015-1025. <https://doi.org/https://doi.org/10.1002/bit.28310>

Wang, S.-J., & Zhong, J.-J. (2007). Chapter 6 - Bioreactor Engineering. In S.-T. Yang (Ed.), *Bioprocessing for Value-Added Products from Renewable Resources* (pp. 131-161). Elsevier. <https://doi.org/https://doi.org/10.1016/B978-044452114-9/50007-4>