fermentation

MDPI

Review

Machine Learning and Deep Learning Strategies for Chinese Hamster Ovary Cell Bioprocess Optimization

Tiffany-Marie D. Baako [1], Sahil Kaushik Kulkarni [1], Jerome L. McClendon [2], Sarah W. Harcum [1] and Jordon Gilmore [1,*]

[1] Department of Bioengineering, Clemson University, Clemson, SC 29634, USA; tbaako@clemson.edu (T.-M.D.B.); kulkar8@clemson.edu (S.K.K.); harcum@clemson.edu (S.W.H.)
[2] Department of Automotive Engineering, Clemson University, Clemson, SC 29634, USA; jmcclen@clemson.edu
* Correspondence: jagilmo@clemson.edu; Tel.: +1-(864)-656-4262

Abstract: The use of machine learning and deep learning has become prominent within various fields of bioprocessing for countless modeling and prediction tasks. Previous reviews have emphasized machine learning applications in various fields of bioprocessing, including biomanufacturing. This comprehensive review highlights many of the different machine learning and multivariate analysis techniques that have been utilized within Chinese hamster ovary cell biomanufacturing, specifically due to their rising significance in the industry. Applications of machine and deep learning within other bioprocessing industries are also briefly discussed.

Keywords: data science; recombinant protein production; deep learning; multivariate statistical analysis; bioprocess engineering; biomanufacturing; Chinese hamster ovary (CHO) cells

Citation: Baako, T.-M.D.; Kulkarni, S.K.; McClendon, J.L.; Harcum, S.W.; Gilmore, J. Machine Learning and Deep Learning Strategies for Chinese Hamster Ovary Cell Bioprocess Optimization. *Fermentation* **2024**, *10*, 234. https://doi.org/10.3390/fermentation10050234

Academic Editor: Massimiliano Fabbricino

Received: 21 March 2024
Revised: 20 April 2024
Accepted: 23 April 2024
Published: 27 April 2024

1. Introduction

Bioprocessing, as with many other manufacturing industries, is at the cusp of an industry-wide shift toward the integration of data-driven approaches for the development of new techniques and the optimization of state-of-the-art approaches. These changes are commonly referred to as Industry 4.0. Industry 4.0 is defined as "the integration of intelligent digital technologies into manufacturing and industrial processes". This includes the integration of cloud computing, the Internet of Things (IoT), data analytics, machine learning, and deep learning technologies with production facilities to optimize manufacturing processes. This paper focuses on how data-driven approaches, including machine learning (ML) and deep learning (DL), are integrated within the biopharmaceutical industry and how different approaches are used to create safe, effective, and efficient biopharmaceutical products. Specifically, this paper focuses on the use and potential use of ML and DL strategies for the bioprocess optimization of Chinese hamster ovary (CHO) cell cultures. This review provides a novel synopsis of the state-of-the-art concerning data-driven approaches and ML/DL for CHO cell bioprocesses, whereas other reviews have documented a more general overview, focusing on multiple cell types and bioprocess applications or more narrowly focusing on specific types of ML/DL [1]. Currently, CHO cell cultures are used for recombinant protein expression, which is one of the main components of drug manufacturing [2]. CHO cells are the industry gold standard due to their capability for post-translational modifications, as well as established standards of good manufacturing practices, compared with other mammalian cells. However, relatively low cell densities, slow growth rate, and low productivity create obstacles for manufacturers due to high demand. These issues affect the entire supply chain and contribute to high manufacturing costs and pricing for these drugs. ML approaches can improve these issues through the implementation of predictive models, data-driven control, and generative artificial intelligence approaches [3].

Within manufacturing processes utilizing CHO cells, fed-batch bioprocesses are the most widely accepted method for feeding and process control due to industrial familiarity and the advantage of incrementally adding in nutrients, buffers, amino acids, and vitamins [4]. These processes have enabled the scalable production of recombinant proteins, monoclonal antibodies, vaccines, and cell-based therapies for various biomedical applications, especially recombinant protein-based pharmaceutical processes of which CHO cell-based products comprise 70% of revenue-generating products [5,6]. The advancement of ML and DL techniques in CHO cell biomanufacturing processes is poised to provide immediate impact within the biopharmaceutical industry, including the improvement of productivity and efficiency and enabling promising new techniques such as continuous biomanufacturing through perfusion cultures.

2. Summary of Common Machine Learning and Deep Learning Approaches

While often used in similar contexts, there exists a difference between machine learning and deep learning. Machine learning (ML) is the capability of a trained machine or model to imitate human behavior and intelligence. Within ML exists the concept of artificial neural networks (ANNs). Data flow through network nodes from the input layer, where the data are fed into the output layer. Deep learning (DL) is a more specific subset of machine learning that involves the use of neural networks with multiple (two or more) hidden layers that derive often unclear connections between inputs and outputs, whereas architectures that consist of a single hidden layer are defined as shallow multi-layer neural networks. These are inspired by the human brain's neural structure. These neural structures comprise multiple layers, which can allow for learning and training based on specific tasks, such as image or speech recognition [2]. DL models are often referred to as "black-box" models because relationships between inputs and outputs made within the hidden layers are either unknown or unintelligible to end users. However, there is significant research focused on implementing explainable AI (XAI) in which ante-hoc or post hoc algorithmic additions are made for black-box models to provide intuitive connections between inputs, model decisions, and outputs [7].

The application of ML and DL within bioprocesses serves as an augmentation to the traditional design of experiments (DoEs) approaches. These approaches have been crucial for optimizing various parameters and improving protein yields. While traditional DoE methods involve systematically changing variables and observing outcomes, ML and DL can complement this by analyzing large amounts of data more comprehensively and identifying patterns or relationships that might not be immediately apparent. Rodriguez-Granrose et al. and colleagues have discussed the integration of DoE with artificial neural network architecture [8].

A DoE approach was used with an artificial neural network for optimizing a bioprocess for cell growth. The system used the design of an experimental approach to find optimal bioprocess set points or, in this case, variables and ANNs were used to model and improve the accuracy of the bioprocess model beyond the capabilities of traditional regression models. The process involved the identification of key variables (cell line, seeding density, media supplement percentage, and media exchange percentage) that have a significant impact on the bioprocess (cell growth). These variables are systematically explored using DoE to determine the most optimal condition for the cell growth process. Since ANNs are trained on experimental data to create a predictive model that has the ability to identify the process more accurately, they are preferred over regression models. This hybrid model outperformed both a standard linear regression model as well as an unoptimized neural network.

Figure 1 represents the optimal ANN that is comprised of 91 functions in a single layer for cell growth modeling. Equal weightage to all four model output functions was provided. As a result, the maximum desirable ANN was determined to be a 1-layer Gaussian ANN. This was also tested in vitro, and the ANN-derived optimal process conditions resulted in 11.2% higher cell doublings as opposed to the linear regression model and 42.2% higher

cell doublings compared to the original "one-factor-at-a-time" experimental condition. This hybrid method of ANN-DOE efficiently leverages the artificial neural network to improve bioprocess outcomes that were previously not possible with linear regression itself [8]. Although this demonstration was not specifically performed on CHO cells, primary cells from the nucleus pulposus tissue of human cadavers, implications for the way that ANN can be leveraged to advance mammalian cell culture productivity are still relevant for CHO-derived recombinant protein bioprocesses. For example, the work of Pinto and coworkers evaluated shallow versus deep neural networks, varying the number of layers between 3 and 5 in a feed-forward neural network (FFNN) using the CHO K-1 cell line [9]. Overall, this work showed that errors were reduced by 14% and 23.6%, respectively, as the number of hidden layers increased, revealing an overall improvement in model accuracy in deep models versus shallow models. However, computational time increased by more than 30%, indicating an important consideration for leveraging ANNs in real-world industrial processes.

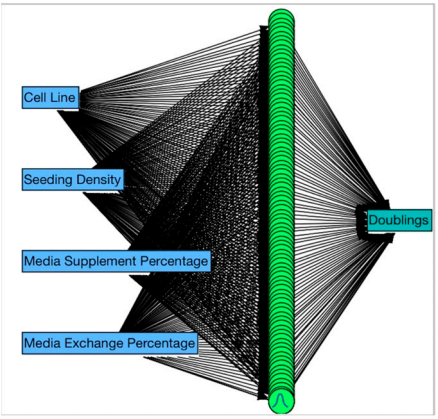


Figure 1. Optimal ANNs comprised of 91 Gaussian activation functions (denoted by the green dots) in a single layer for cell growth modeling [8]. Reprinted with permission from the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). Accessed on 1 April 2024.

2.1. Machine Learning in Biopharmaceutical Manufacturing

Within ML, two major paradigms exist for model-based approaches and supervised and unsupervised learning. Supervised learning (SL) is a class wherein the models are trained on labeled datasets, where the algorithm learns to map input features to known output labels [10]. In the context of Chinese hamster ovary cells, various models relating to neural networks have been deployed for predictive modeling. For instance, Hisada et al. utilized the LASSO (Least Absolute Shrinkage and Shrinkage Operator) model, which specialized in predictive score metrics and anomaly detection to predict monoclonal antibody (mAb) productivity changes in CHO cells based on morphological profiles [11]. They constructed the LASSO model using 368 morphological parameters as explanatory variables to achieve high accuracy in predicting mAb productivity simply from morphological profiles [11]. This work developed a methodology to predict changes in antibody production in CHO cells using morphological profiling techniques. The researchers optimized an image acquisition pipeline to capture stable morphological profiles of suspension cells and utilized specific morphological parameters for anomaly prediction calculations. By combining morphological profiling with machine learning, the researchers successfully predicted mAb productivity changes solely from early morphological profiles, demonstrating the potential for the non-invasive and quantitative evaluation of subtle quality changes in host cells [11].

In the context of bioprocess engineering relating to specific cell culture types, various forms of machine learning have been used for predicting outcomes based on historical data, such as optimizing nutrient concentrations for maximum cell growth, cell density, and pro-

tein yield [12]. Unsupervised learning (UL) involves exploring various data patterns without labeled information. These techniques are often applied to clustering similar batches or samples for the identification of process variable anomalies or the self-organization of data, thus assisting qualitative control mechanisms [12].

From a more fundamental standpoint, most ML algorithms can be broken down into two categories: deterministic or probabilistic. Deterministic models are usually well-defined, rule-based algorithms that provide the same model output each time for a particular set of inputs. One example of a Boolean-based deterministic model is the decision tree [13–15], which has been successfully applied in the optimization of the fermentation process and the identification of fermentation parameters. However, the downside of deterministic models is that they do not account for the inherent variability that may be in a system. In Kumar et al.'s work, regression models have been employed to analyze and optimize bioprocessing unit operations [16].

2.2. Deep Neural Networks in Biopharmaceutical Manufacturing

Neural networks consist of interconnected nodes which process information in different layers. Neural networks are inspired by the human brain, which consists of more than several billion neurons that segregate into three parts: dendrites, soma, and axons [17]. The primary function of these three components is to receive, transmit, and connect information from one neuron to another.

In the context of biopharmaceutical manufacturing, deep neural networks are used for optimal glycosylation analysis in CHO cells. For instance, Kotidis et al. and colleagues utilized deep neural networks to predict the glycosylation outcomes of monoclonal antibodies produced in CHO cells. By training the ANN model with experimental data on intracellular nucleotide sugar dynamics and extracellular metabolite concentrations, they accurately predicted site-specific glycoform distributions. The ANN model successfully captured the effects of metabolic perturbations, manganese supplementation, and glycosyltransferase knockouts on glycosylation outcomes, showcasing its potential for optimizing glycosylation processes in CHO cell production. Neural networks were crucial in this study for accurately predicting site-specific glycoform distributions of recombinant glycoproteins in CHO cells, offering a data-driven approach that requires minimal biological background knowledge and enables rapid model development for optimizing glycosylation processes, thus enhancing manufacturing efficiency [18].

Similarly, Antonakoudis et al. developed a hybrid modeling framework that combines mechanistic information in the form of a stoichiometric model with a deep artificial neural network to predict antibody glycosylation patterns in Chinese hamster ovary cells [19]. By training the neural network with data on bioprocess variables such as metabolite fluxes, cell culture parameters, antibody quality parameters, and glycosylation pathways, the researchers leveraged deep neural network principles to enhance the prediction of product quality (glycan distribution) in bioprocesses involving CHO cells. Within the biopharmaceutical context, these networks can be used for various tasks, such as process optimization based on feed data, quality prediction based on previous attributes, and overall process efficiency due to their predictive and qualitative nature [20].

2.2.1. Recurrent Neural Networks

Recurrent neural networks (RNNs) are used to model temporal dependencies, such as biomass concentration, pH, and temperature in bioprocess data, which capture sequential patterns such as fermentation dynamics. They have the ability to play a significant role in optimizing biopharmaceutical upstream processes by providing reliable estimates for key process parameters and enabling the study of growth and metabolite-related outcomes over time. RNNs are also used for the prediction of process outcomes and the control of upstream processes. By leveraging historical data and learning patterns from past observations, RNNs can forecast future processing behavior and help make informed decisions for process optimization.

For instance, Smiatek et al. and colleagues harnessed the power of recurrent neural networks (RNNs) to effectively model and predict the behavior of CHO cells in biopharmaceutical upstream processes. Through the development of specific and generic RNN models, they have been able to analyze complex temporal data patterns associated with CHO cell growth, metabolite concentrations, and product titers. By leveraging the capabilities of RNNs, the researchers gained valuable insights into how different process conditions impact CHO cell performance, enabling informed decision-making for process optimization in biopharmaceutical manufacturing. Recurrent neural networks are crucial in this study for their ability to effectively model and predict the dynamic temporal sequences of key process parameters in biopharmaceutical upstream processes involving CHO cells [21]. The high predictive accuracy and flexibility of RNN models have proven instrumental in capturing the dynamic nature of CHO cell-based bioprocesses and facilitating advancements in biopharmaceutical development. As these variables are classified as temporal dependencies, RNNs are suitable for capturing the sequential nature of these variables and can be trained to predict future anomalies. This is pivotal for the prediction and optimization of time-dependent variables in bioprocessing [22].

2.2.2. Convolutional Neural Networks

A convolutional neural network (CNN) is a type of neural network that is designed specifically for processing structured grid-like data, such as images or sequences. It has multiple layers, including convolutional layers. Convolutional neural networks excel in extracting spatial features, making them valuable for image-based bioprocess monitoring, and can be applied to analyze the microscopic images of cell cultures, which assists in identifying irregularities and optimizing growth conditions [23].

For example, Wang et al. utilized the concept of in situ microscopic imaging and deep learning-based image analysis to monitor and analyze the culture process of Chinese hamster ovary (CHO) cells [24]. They employed a deep learning-based Mask R-CNN algorithm for the segmentation and analysis of online pre-collected images from the suspension culture of CHO cells. The researchers used three different methods of data augmentation and transfer learning technology to improve the performance of the model.

As depicted in Figure 2, CHO cells have been cultured inside a bioreactor along with a vision probe. The monitor displays the collected images together with the image analysis result. The in situ microscope, equipped with an imaging system and connected to a computer, captures cell images in real-time during the cell culture process. These images are then subjected to deep learning-based image analysis using the Mask R-CNN algorithm, allowing for the accurate segmentation and analysis of the cells. The deep learning method, trained by 183,040 labeled cells of 184 images, effectively-recognized cells within clusters. This approach outperforms traditional methods in terms of common image-based evaluation methods. The researchers analyzed the temporal variations in cell dimensions and shapes, indicating that the majority of cells are approximately spherical, with their average diameter increasing throughout the culture process [24].

Convolutional neural networks (CNNs) have been specifically used in the context of CHO cells to train the Mask R-CNN algorithm for cell segmentation and analysis. This deep learning-based approach has demonstrated promising performance in accurately identifying and quantifying cell characteristics, providing valuable insights into the culture process of CHO cells. Apart from cell culture image analysis, CNNs can be trained to analyze these microscopic images to assess the viability and metabolic activity of microbial cells during fermentation processes, contamination detection, yield prediction, and image-based scale-up optimization [25].

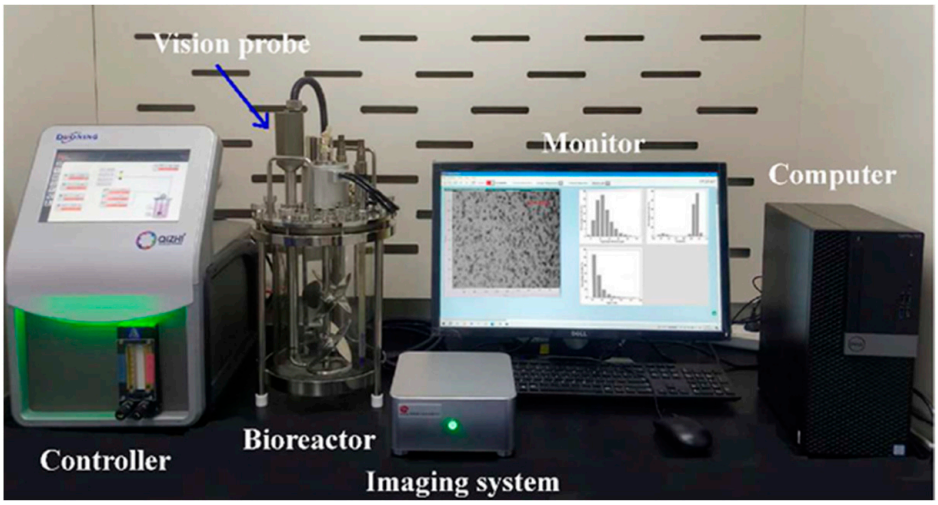


Figure 2. Platform of ongoing cell culture with vision probe [24]. Reprinted with permission from Elsevier.

2.2.3. Generative Artificial Intelligence

Generative models are a subset of machine learning models that are specifically designed to generate new, synthetic data that resemble a given set of training examples. Unlike discriminative models, which focus on classifying or predicting labels, generative models learn the underlying distribution of the training data and can generate new samples from that distribution [6]. One such type is the generative adversarial networks (GANs) comprised of a generator and a discriminator neural network. While the discriminator attempts to distinguish between real and generated samples. The generator attempts to create realistic data. With the help of adversarial training, GANs improve over time, thus generating increasingly convincing synthetic data. cGANs, otherwise known as conditional generative adversarial networks, are one step ahead by introducing a conditioning variable [26]. By associating the conditioning variable with factors like loading or the environment, cGANs are valuable for modeling different structural behaviors. The cGANs learn and replicate the data's underlying distribution, enabling applications in data augmentation and uncovering complex relationships within structures. In bioprocessing, generative neural networks have been employed for generating synthetic data that can mimic the characteristics of real bioprocess data, which helps in augmenting limited datasets and facilitates model training in diverse situations [27].

3. Bioprocess Data and ML/DL Targets

Bioprocess data are typically collected within a time series, where the targeted data are collected at predetermined time increments, with the intention of using these collected values to inform process control, predict future values, or detect anomalies within the process. The parameters that are typically monitored during bioprocesses include several online and offline parameters. Online parameters include controlled variables, such as pH, dissolved oxygen (DO), and the media and feed rates in fed-batch and continuous processes. These online variables can be manipulated by cellular processes (waste generation, apoptosis, growth, replication, etc.) but also by control variables such as base addition (controls pH), stir speed (controls DO), and gas flow (which controls/balances off-gas measurements). Offline variables are taken from samples throughout the culture process and include parameters such as viable cell density (VCD), metabolite concentrations, and protein titer. Several commercial instruments are available to measure these parameters, such as the Cedex BioAnalyzer (Roche; Basel, Switzerland), Octet (ForteBio; Fremont, CA, USA), or BioProfile® FLEX2 (NOVA Medical; Waltham, MA, USA).

Within biomanufacturing, there are three approaches that are utilized to produce biopharmaceuticals and other products as follows: batch, fed-batch, and continuous ap-

proaches. The mode of the process is chosen based on factors such as the needs of the host organism, efficiency concerns, and product demand. In a batch bioprocess, the process is conducted in a single batch; thus, all the feed and other substances that are necessary for product production are added at the beginning of the batch, and then the product is harvested once the process is complete. These processes are simple in comparison to the other modes; however, they are not as commonly performed due to the extensive time needed to complete a batch, as well as the frequent need to repeat batches to acquire the desired product yield. Fed-batch bioprocessing is a derivation of batch processes in that raw materials and other substances are added to the bioreactor at the beginning of the process, but additional nutrients (including feeds and supplements) are also added during the process. This method is an improvement from batch processes because it allows greater control of cell growth and product yield. The continuous method of biomanufacturing is denoted by the lack of distinct batches within a process. The raw materials and needed supplements are systematically added until the process is manually ended. This method is typically chosen when there is high demand for the product due to its high throughput nature. One of the major benefits of continuous processes that have led many researchers to push for it to become the standard of biomanufacturing is that it can reduce manufacturing costs due to its capability of using more compact laboratories and other facilities [28,29]. Another advantage of continuous biomanufacturing is that it provides improved quality control. For example, Bayer has developed continuous bioreactor technology that minimizes the product residence time such that it does not remain in non-ideal conditions (such as temperatures where degrading enzymes are more active) for long periods of time [30]. Although this methodology shows promise, there are still certain drawbacks that have prevented it from becoming more mainstream in the biopharmaceutical industry. One of the most glaring issues is the need for cell retention devices; these can add operational complexity [31]. Cell densities are significantly higher within continuous processes; thus, the cells require particular attention to prevent the culture from depleting resources [31], and there is a risk of CO_2 and other byproduct accumulation [28]. Without proper attention to process controls, the system can become overwhelmed and lead to failed cultures.

Online process control parameters are typically set at the beginning of an experiment but can be adjusted throughout the process according to necessities that may arise due to cultural behavior. The typical process setting ranges for CHO cell bioreactors for temperature, DO, and pH at 34–37 °C, 30–50%, and 6.7–7.3, respectively [32]. These process settings are widely used due to their similarity to human and Chinese hamster body conditions. To control these parameters to the setpoints, bioreactors typically use proportion-integral-derivative (PID) control loops. It has been shown that PID settings can also affect culture outcomes [33].

Due to the stirring and gas sparging within bioreactors during a process to control DO, antifoam is commonly added to control foaming. Yet, it is important to note that antifoam additions are used sparingly as these chemicals can reduce production and cell growth [34]. Another critical online parameter for fed-batch and continuous bioprocesses is the feed profile. Feeds are added to cultures to prevent nutrient depletion and increase product titers by extending the culture durations. These feeds are usually more concentrated than the basal media and contain nutrients such as glucose, vitamins, and amino acids to increase the productivity of a culture [35].

Several offline parameters are commonly measured to gauge the culture's health. Additionally, product titers are often obtained offline. As the product of CHO cell cultures are normally large biomolecules such as monoclonal antibodies, the glycosylation and charge variant state of the molecules may also be obtained. These properties of the protein are known as critical quality attributes (CQAs). Protein titer refers to the concentration of a particular protein of interest. Monoclonal antibodies (mAbs) are immunoglobulins that are of particular interest as drugs due to the high specificity of these molecules toward receptors in the human body. There are currently 92 mAbs approved by the Food and

Drug Administration (FDA) and European Medicine Agency (EMA) for the treatment of autoimmune diseases and cancers in 2022 [36].

Glycosylation represents a wide range of carbohydrate moieties added to a protein at specific amino acids, also known as post-translational modification. These glycan additions affect the protein function in a human patient [37] and long-term efficacy [38]. As of 2018, 62 of the 71 new biopharmaceutical active ingredients that were introduced to the market were recombinant proteins, and 52 of those were derived from mammalian cell lines due to high post-translational modification capability [39]. Due to the high potential of these mAbs for therapeutics, there have been many advances made in ML to increase mAb protein production.

Another key process parameter is viable cell density (VCD, cells per mL of the culture broth), which has become yet another prime target for optimization in ML and DL applications due to it being a key parameter in determining the success of a culture.

Table 1 summarizes typical online and offline parameters that are monitored during cell culture processes.

Table 1. Common online and offline parameters in biomanufacturing.

Online Parameters	Offline Parameters
Temperature	Protein titer
pH	VCD
Base additions	Metabolite concentrations (Glc, Lac, Gln, Glu, NH_3)
Dissolved oxygen (DO)	Glycosylation
Feed rates	Amino acid concentrations
Vessel Pressure	Osmolarity
Off-gas concentrations	Charge variants
Stir speed	
Antifoam additions	

Abbreviations: viable cell density (VCD), glucose (Glc), lactate (Lac), glutamine (Gln), glutamate (Glu), and ammonia (NH_3).

The online data obtained from bioreactor cultures can be expansive, with upwards of 15,000 data points for each input parameter/feature. While this may lead one to believe that there is no lack of data available for training ML and DL models, this is one of the current bottlenecks that researchers are working to eliminate. While the quantity of data is expansive, they lack in quality. A very large fraction of the data available from CHO cell cultures is largely homogeneous due to similar media and feed formulations being used, as well as other process controls. This creates an issue of overfitting when training predictive models for optimizing cultures. When the models are trained on data that do not capture the entirety of cell behavior within a culture with varying conditions, this can cause the model to become highly accurate when predicting the values, such as VCD and titer, as long as the input variables are similar to those in the dataset used for training. If the features are outside the scope of these data, then the model's predictive accuracy greatly decreases. Thus, there is a need to have a diversity of model input data for improved model robustness across all possible culture conditions.

With large amounts of bioprocess data comes the need for pre-processing prior to their use for modeling. This pre-processing includes data curating, feature selection, transformation, and reduction. Curating typically entails imputation for missing values, outlier detection, and treatment or removal, and noise reduction using methods such as signal processing or smoothing. Feature selection entails selecting process variables based on their relevance to the modeling or prediction task, and irrelevant or redundant features are removed from the data [40]. After the features are selected, they are then transformed to improve ease of implementation. This can include but is not limited to aggregating the data over time intervals, normalizing and standardizing the data using methods such as k-nearest neighbors and support vector machines, and categorizing variables [41,42].

The treated data can then be split into training, validation, and testing sets. Once this is complete, steps can be taken to reduce the dimensionality of the data using methods such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) [43]. The parameters can then be encoded into numerical representations using methods available through libraries such as Scikit-Learn so that they may be easily fed into ML models [44].

4. Machine Learning Approaches for Critical Quality Attribute Optimization

The primary goal for using ML in biomanufacturing is to improve product yield (titer) and quality. VCD is a secondary goal, as the product titer is highly correlated with VCD. While statistical models may be used to correlate the relationship between CQAs and critical process variables (CPPs), mechanistic hybrid ML models have increased in prevalence due to their ability to describe the causation in cultural behavior [45]. One such demonstration of this method in predicting VCD is the development of a hybrid agent-based approach where the CHO cells were treated as individual agents with a flux balance model that predicts changes in metabolite and nutrient concentrations [46]. The initial cell culture conditions and measured DO and sodium levels were used as the model inputs, and consequently, the model was able to predict VCD, among other CQAs. When the initial VCD was low, the root mean squared error (RMSE) was 2.48; however, when the initial VCD was high, the RMSE was 4.44. Despite this, the hybrid agent-based model was increasingly accurate in predicting glucose and lactate, which is a byproduct that can inhibit cell growth levels [46]. Another previous work utilizing a hybrid model for VCD prediction combined a multilayer perceptron (MLP) regressor, or feed-forward neural network, random forest (RF), and extreme gradient boosting (XGBoost) algorithm (all developed using Python 3.9.12 default libraries) with mechanistic equations to predict next-day VCD and other CQAs [47]. Inputs for the ML algorithms were first separated by which specific calculated rate they applied to (specific growth rate, specific productivity, or specific cumulative glucose consumption rate). The correlation analysis and multicollinearity assessment were then conducted to determine the optimal combination of inputs based on process knowledge and statistical analyses. Hyperparameters for the ML algorithms were determined based on prior knowledge from previous works. Among the selected hyperparameters for the RF regressor were the number of estimators, maximum depth, and minimum split, while the number of estimators, learning rate, maximum depth, and subsample ratio were optimized for the XGBoost regressor. The number of neurons, activation function, and solver were all adjusted to optimize the MLP regressor. Once the VCD was normalized after considering all time points in all the experiments that were conducted, the MLP had the highest RMSE but the lowest mean absolute error (MAE); meanwhile, the XGBoost had the lowest RMSE but the highest MAE. It was also discovered that all three models were capable of accurately predicting VCD for a second CHO cell line. Advances in VCD prediction, such as those previously discussed, have provided a means to enhance therapeutic protein production.

Glycosylation has become a target CQA for process modeling and optimization because of its ability to decrease therapeutic protein immunogenicity through increased protein solubility and stability [48]. Kinetic models have previously been used due to their means of efficiently demonstrating the cellular mechanisms that comprise glycosylation. However, they are not capable of accurately predicting the specific sites that the glycan moieties take on without large quantities of kinetic variables and an understanding of the enzyme and protein levels within the cells, as well as extensive time for training [18,49]. Therefore, machine learning approaches have been developed to address these pitfalls. Kotidis and Kontoravdi developed an ANN to predict the site-specific glycoform distributions of four recombinant proteins that are expressed in three CHO cell lines (GS-CHO, CHO-K1, and CHO-S), two IgG monoclonal antibodies, and two fusion proteins [18]. They observed that the ANN, both on its own and as part of a hybrid model that paired CHO metabolism kinetics with the data-driven model, outperformed the standalone kinetic

model and was successful in modeling and predicting the glycoform distributions with average absolute errors as low as 0.98%. Another approach that has been taken in many previous studies is utilizing data-driven models such as PCA and partial least squares regression (PLSR) to understand the variables that, in turn, affect the glycosylation levels of a culture and how they correlate with one another [50]. For example, Powers et al. were able to determine how different types of media resulted in different terminal galactosylation rates, how high mannose levels were not correlated with aglycosylation, and how initial glucose concentrations were positively correlated with galactosylation rates using PCA and other multivariate data analysis (MVDA) methods [51]. The use of PLSR also enables the ease of feature selection. Variable importance in projection (VIP) is a method of feature selection that effectively quantifies the effect that process variables have on CQAs. VIP is also useful in preventing overfitting, as demonstrated when compared to a previously developed genetic method in the context of glycan value predictions using PLSR [52]. The utilization of the aforementioned methods in the prediction of glycosylation patterns also has the potential to increase understanding of other post-translational modifications that influence potency and immunogenicity.

As previously stated, the demand for therapeutic proteins is ever-increasing, thus creating the need for high protein yields in cultures. There have been various approaches taken to achieve this, including predictive modeling, pattern recognition, and real-time monitoring. For example, Le et al. utilized support vector regression (SVR) and PLSR to predict final protein titer and lactate concentration [53]. They observed that the models had similar high accuracies when the bioreactor production data they were trained on utilized either the final titer or final lactate concentration as the objective function. When the data that were collected during the inoculum train were solely used to train the models, the predictive accuracies for the final titer decreased. However, the seed train data proved to be useful in predicting which bioreactor run would be highly or lowly productive. They found that the critical variables in both the inoculum train stage and during production were associated with lactate metabolism and cell growth and determined that low specific glucose consumption was correlated with low lactate production and consumption. From these findings, it was concluded that corrective measures to exploit the CHO cell metabolism to increase productivity while limiting lactate production would need to occur within the first 70 h of the run. An example of employing real-time monitoring to optimize titers is the use of ML to predict the final protein concentrations in bioreactors. Tulsyan et al. developed strategies that comprised switching ML algorithms in real-time for non-linear state estimation [54] and adaptive state simulation [55]. Bayrak et al. capitalized on these previous works and developed a methodology in which SVM, PLSR, Gaussian process regression (GPR), regression trees (RT), and ensemble trees (ET) are each evaluated on their performance in learning from real-time data, and the highest performing model is chosen in real-time for titer prediction for the proceeding day [56]. It was observed that utilizing dynamic feature selection substantially reduced the root mean squared error of cross-validation (RMSEcv) for the PLSR, GPR, and SVM models; thus, it was implemented for the continuation of the study. At the conclusion of the study, it was observed that the PLSR and GP algorithms maintained the lowest error at each time step, thus proving that they are strong candidates for real-time process monitoring [56].

Table 2 includes a summary of the aforementioned ML and multivariate techniques that have been utilized for CQA optimization.

Table 2. Summary of ML and MVDA applications in CQA optimization.

Target	Approach	Reference
VCD prediction	Hybrid agent-based, flux balance model	[46]
	MLP, RF, XGBoost	[47]
Glycosylation prediction	ANN, metabolic kinetic model	[18]
	PCA, PLSR	[51]
	PLSR	[52]
Protein titer prediction	SVR, PLSR	[53]
	SVM, PLSR, GPR, RT, ET	[56]

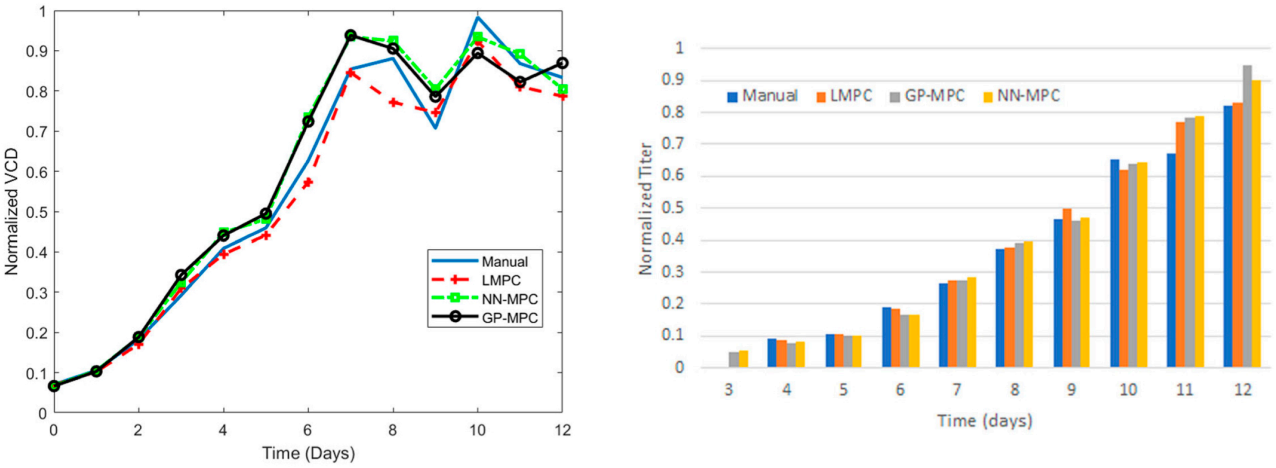
Abbreviations: viable cell density (VCD); multilayer perceptron (MLP); random forest (RF); extreme gradient boost (XGBoost); artificial neural network (ANN); principal component analysis (PCA); partial least squares regression (PLSR); support vector regression (SVR); support vector machine (SVM); Gaussian process regression (GPR).

5. Machine Learning Approaches for Process Control

As previously mentioned, an important aspect of bioreactor process control is PID control. Bioprocesses can be highly variable, non-linear, and complex, thus making it difficult to control various variables to optimize CQAs and product production. Therefore, it has become common practice to utilize PID control, a continuous closed-loop feedback control. This approach uses a sensor that measures and compares various process variables against their setpoint values and then sets the difference as the parameter error. The PID algorithm then adjusts the control output to decrease the parameter error [57]. There are three variables within the algorithm that determine the controller output: proportional gain (k_p), integral gain (k_i), and derivative gain (k_d). The proportional gain term influences the controller output to be proportional to the system error at each timestep. k_i is dependent on the combined previous errors and how far they deviate from the variable setpoint. Derivative gain is determined by the rate of change in the controller output based on the parameter error it receives over time. The modification of these three variables consists of carefully defining each term based on the nature of the process. It has been demonstrated that utilizing this methodology for PID tuning can have a positive influence on heating the regression rate and setpoint control during prolonged periods of liquid addition within a bioreactor system [58], thus making it an integral optimization target in biomanufacturing.

Another design approach that has become more prominent within biomanufacturing is iterative modeling. This process typically consists of designing an experiment based on key variables of interest, collecting the experimental data, modeling them via computational and/or physical models, and then making design adjustments based on the results after they have been validated. Park et al. demonstrated this by creating a systematic media design framework that consisted of culture data collection, multivariate statistical analysis, in silico flux analysis with a genome-scale metabolic model, and knowledge-based targeting media components to determine the optimal media and feed combination to increase IgG1 yield and eliminate cellular bottlenecks that impact cell growth and production [59]. Through this methodology, they were able to determine that one media and feed combination out of the four that were studied outperformed the others. They then analyzed the metabolic fluxes of the CHO cells in the optimal media/feed combination and were able to conclude that an aspartate supplement could enhance growth and productivity within the culture. Another utilization of this method is in the development of data-driven model predictive controls (MPCs) to understand the dynamics between glucose concentration and CQAs such as VCD, osmolality, and byproduct production [60]. MPCs were derived based on linear regression, Gaussian process regression (GPR), and neural networks and were coupled with a feed–glucose mass balance relationship in order to simulate the fed-batch process. After comparing these MPCs to a rule-based control technique, they found that the GPR-derived MPC outperformed the other models in maximizing VCD (Figure 3a) and the titer (Figure 3b), with the neural network-derived MPC following closely after. At the conclusion of the experiment, it was determined that utilizing the non-linear MPCs on volume-based datasets could increase protein production by over 10% [60]. These

demonstrations of iterative design are a testament to the promise that this design approach has for biomanufacturing optimization.



(**a**) (**b**)

Figure 3. Trajectories for GPR-MPC, NN-MPC, linear-MPC, and manually controlled culture predictions for (**a**) viable cell density and (**b**) protein titer [60]. Reprinted with permission from John Wiley and Sons.

6. Emerging Topics beyond CHO Cell Bioprocesses

6.1. Cell Types in Bioprocesses for Machine Learning and Deep Learning

Mammalian cells are optimal for bioprocess engineering as they excel in protein folding and post-translational modifications. Their property of replicating human glycosylation patterns is ideal for the creation of biopharmaceutical products and has lower chances of causing immune reactions [61]. Bacterial cells are also commonly used in bioprocess engineering due to their genetic makeup, rapid growth, and simple cultivation methods, which are all qualities that allow for the production of proteins, enzymes, and other compounds on a scale. Their ease of modification and adaptability to environments make them versatile and cost-efficient bioprocess cell sources [62]. Along with mammalian and bacterial cells, viruses are extensively used in bioprocessing because of their ability to effectively transport genetic material into host cells, which is essential for gene therapy and viral-based vaccines. Their ability to transfer genes to targets through viral attachment and entry methods enables precise control over cellular activities [63]. Additionally, insect cells, while being very similar to mammalian cell sources, can be cultured at higher densities, exhibiting simpler post-translational modification patterns compared to mammalian cells, and are less susceptible to contamination by mammalian viruses. This makes insect cells safer for certain types of bioprocessing [64].

Bacterial cells, for example, *Escherichia coli* (*E. coli*), are prokaryotic microorganisms that can be characterized by their lack of a membrane-bound nucleus and other organelles, with genetic material organized in a single circular chromosome [65]. As previously mentioned, bacterial cells' main strength is their ease of modification and adaptability. This enables bacterial cells to produce several biopharmaceutical products, including biofuels such as ethanol, organic acids such as lactic, succinic acid, and polyhydroxyalkanoates (PHAs). These cells are engineered to efficiently convert various substrates into recombinant proteins and mAbs. The ability of bacterial cells to grow rapidly, be genetically modified, and exhibit low contamination risks makes them an ideal candidate as a cell type source for bioprocess engineering strategies [66]. Further advancing biomanufacturing with these bacteria cells, researchers have sought to combine machine learning with metabolic models. One such example is the work of Oyetunde and co-workers, who utilized support vector machines, gradient-boosted decision trees, and neural networks in a stacked regressor model to build

a predictive model for engineered *E. Coli*, producing recombinant proteins. This approach appeared to be largely successful, with a correlation coefficient of 0.93 on the validation dataset [67].

Viruses, for example, baculovirus, as intracellular parasites, rely on host cells for replication. These are often used as delivery vectors to introduce mRNA molecules into host cells. These mRNA molecules encode the proteins. Once the mRNA molecules are inside of the host cell, the mRNA is translated by the cellular mechanisms to produce the corresponding proteins [68]. Viral vectors, such as adeno-associated viruses (AAVs) or lentiviruses, are engineered to efficiently deliver mRNA into target cells, ensuring high levels of protein expression. This methodology enables the production of recombinant proteins or antigens in a controlled and scalable manner, making it a promising strategy for upstream manufacturing, including mRNA-based vaccine development and gene therapy [69]. One challenge to the advancement of AAV-based therapies is the selection of the proper mRNA targets for translation. This limiting factor is addressed through machine learning-enabled synthetic biology approaches to the design of viral vectors [70,71].

Insect cells are utilized as one of the ideal cell types in bioprocess engineering due to their capacity for performing complex post-translational modifications similar to mammalian cells, robust scalability in suspension culture systems, and general flexibility in expressing a wide range of recombinant proteins, collectively enabling the efficient large-scale production of biopharmaceuticals and viral vectors [72]. In the research conducted by Marwidi et al. and colleagues, insect cell lines were used for the manufacturing of AAV vector production. The insect cell line used was Sf9, which is a common cell line that is used in baculovirus-insect cell systems for AAV vector production. This ability of insect cells to produce viral vectors is transferable to bioprocess engineering and upstream manufacturing applications as well. For example, scalability, gene expression control, and quality control are valid reasons to consider insect cell lines, specifically Sf9, for manufacturing purposes. The ability to modulate the expression level and content and individual Rep and Cap proteins is an ideal example of precise gene expression control in bioprocess engineering for AAV vector production [73]. Machine learning can be used to further capitalize on the scalability of insect-cell-based systems, such as in the work of Altenburg et al., which featured double differential digital holographic microscopy and machine learning to determine the precise moments in which viability and cell density were optimized [74]. This information could then be used to determine strategies for replicating or prolonging these optimized culture conditions.

6.2. The Rising Need for Explainable Artificial Intelligence

A topic that has become increasingly prevalent within AI and ML is explainable artificial intelligence (XAI), which refers to the ability of AI systems to provide justification and explanations for the decisions or outputs that are provided by many "black-box" models in a manner that is understandable to users. Moving forward, XAIs will be critical as black-box models are used more in bioprocess optimization. There are various terms used to denote the level of interpretability that AI models possess, such as transparent or opaque, model-agnostic or model-specific, explanation by simplification or feature relevance, and visual or local explanation [75]. The National Institute of Standards and Technology (NIST) recognizes four fundamental principles that determine whether a form of AI is to be considered XAI or not: explanation, meaningfulness, accuracy, and knowledge limits [76]. These standards are set forth to promote the societal acceptance of AI utilization, emphasize the safe usage of AI, and increase knowledge of the policy regarding it, all of which are critical for numerous industries, including biomanufacturing. The implementation of XAI in biomanufacturing can improve the state of the art by contributing to factors such as process optimization and quality assurance, as well as providing the necessary transparency needed in order to train operators on the skills needed for bioprocess control and inform consumers and stakeholders of the role that AI plays in biomanufacturing.

7. Conclusions

This review addresses various machine learning and deep learning strategies for data-driven bioprocess engineering. The core of Industry 4.0 is to converge digital technologies, automation, and powerful data to transform, optimize, and ease traditional industrial processes. A wide range of machine learning and deep learning methodologies are being discussed, including the examination of various models, mathematical approaches, and the significance of mathematical equations for bioprocess engineering. As this field continues to evolve, the integration of ML and DL approaches in the biotechnology/pharmaceutical industry promises to revolutionize the standards of innovation without compromising on quality, safety, and data integrity.

Author Contributions: The following list describes the authorship contributions for this review. Conceptualization of paper structure and content was conducted by T.-M.D.B. and J.G. Original draft preparation (writing) was conducted by T.-M.D.B. and S.K.K. All authors reviewed and edited the document, with J.G. leading this effort. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC was funded by the Advanced Mammalian Biomanufacturing Innovation Center (AMBIC), which is an Industry–University Cooperative Research Center Program under the US National Science Foundation (grant no. 1624684).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khanal, S.K.; Tarafdar, A.; You, S. Artificial intelligence and machine learning for smart bioprocesses. *Bioresour. Technol.* **2023**, *375*, 128826. [CrossRef] [PubMed]
2. Yang, C.-T.; Kristiani, E.; Leong, Y.K.; Chang, J.-S. Big data and machine learning driven bioprocessing—Recent trends and critical analysis. *Bioresour. Technol.* **2023**, *372*, 128625. [CrossRef]
3. Duong-Trung, N.; Born, S.; Kim, J.W.; Schermeyer, M.-T.; Paulick, K.; Borisyak, M.; Cruz-Bournazou, M.N.; Werner, T.; Scholz, R.; Schmidt-Thieme, L.; et al. When bioprocess engineering meets machine learning: A survey from the perspective of automated bioprocess development. *Biochem. Eng. J.* **2023**, *190*, 108764. [CrossRef]
4. Chen, C.; Wong, H.E.; Goudar, C.T. Upstream process intensification and continuous manufacturing. *Curr. Opin. Chem. Eng.* **2018**, *22*, 191–198. [CrossRef]
5. Yee, J.C.; Rehmann, M.S.; Yao, G.; Sowa, S.W.; Aron, K.L.; Tian, J.; Borys, M.C.; Li, Z.J. Advances in process control strategies for mammalian fed-batch cultures. *Curr. Opin. Chem. Eng.* **2018**, *22*, 34–41. [CrossRef]
6. Pereira, S.; Kildegaard, H.F.; Andersen, M.R. Impact of CHO Metabolism on Cell Growth and Protein Production: An Overview of Toxic and Inhibiting Metabolites and Nutrients. *Biotechnol. J.* **2018**, *13*, e1700499. [CrossRef]
7. Hu, Y.; Qin, L.; Li, S.; Li, X.; Zhou, R.; Li, Y.; Sheng, W. Adaptive corrected parameters algorithm applied in cooling load prediction based on black-box model: A case study for subway station. *Energy Build.* **2023**, *297*, 113429. [CrossRef]
8. Rodriguez-Granrose, D.; Jones, A.; Loftus, H.; Tandeski, T.; Heaton, W.; Foley, K.T.; Silverman, L. Design of experiment (DOE) applied to artificial neural network architecture enables rapid bioprocess improvement. *Bioprocess Biosyst. Eng.* **2021**, *44*, 1301–1308. [CrossRef] [PubMed]
9. Pinto, J.; Ramos, J.R.C.; Costa, R.S.; Rossell, S.; Dumas, P.; Oliveira, R. Hybrid deep modeling of a CHO-K1 fed-batch process: Combining first-principles with deep neural networks. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1237963. [CrossRef]
10. Chaouch, S.; Yvonnet, J. An unsupervised machine learning approach to reduce nonlinear FE2 multiscale calculations using macro clustering. *Finite Elements Anal. Des.* **2024**, *229*, 104069. [CrossRef]
11. Hisada, T.; Imai, Y.; Takemoto, Y.; Kanie, K.; Kato, R. Prediction of antibody production performance change in Chinese hamster ovary cells using morphological profiling. *J. Biosci. Bioeng.* **2024**, *in press*. [CrossRef]
12. Allenbrand, C. Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews. *Healthc. Anal.* **2023**, *5*, 100288. [CrossRef]
13. Yang, Y.; Farid, S.S.; Thornhill, N.F. Data mining for rapid prediction of facility fit and debottlenecking of biomanufacturing facilities. *J. Biotechnol.* **2014**, *179*, 17–25. [CrossRef] [PubMed]

14. Buck, K.K.S.; Subramanian, V.; Block, D.E. Identification of Critical Batch Operating Parameters in Fed-Batch Recombinant *E. coli* Fermentations Using Decision Tree Analysis. *Biotechnol. Prog.* **2002**, *18*, 1366–1376. [CrossRef] [PubMed]
15. Coleman, M.C.; Buck, K.K.S.; Block, D.E. An integrated approach to optimization of *Escherichia coli* fermentations using historical data. *Biotechnol. Bioeng.* **2003**, *84*, 274–285. [CrossRef] [PubMed]
16. Kumar, V.; Bhalla, A.; Rathore, A.S. Design of experiments applications in bioprocessing: Concepts and approach. *Biotechnol. Prog.* **2013**, *30*, 86–99. [CrossRef] [PubMed]
17. Kakkar, S.; Kwapinski, W.; Howard, C.A.; Kumar, K.V. Deep neural networks in chemical engineering classrooms to accurately model adsorption equilibrium data. *Educ. Chem. Eng.* **2021**, *36*, 115–127. [CrossRef]
18. Kotidis, P.; Kontoravdi, C. Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab. Eng. Commun.* **2020**, *10*, e00131. [CrossRef]
19. Antonakoudis, A.; Strain, B.; Barbosa, R.; del Val, I.J.; Kontoravdi, C. Synergising stoichiometric modelling with artificial neural networks to predict antibody glycosylation patterns in Chinese hamster ovary cells. *Comput. Chem. Eng.* **2021**, *154*, 107471. [CrossRef]
20. Mahdiraji, H.A.; Yaftiyan, F.; Abbasi-Kamardi, A.; Garza-Reyes, J.A. Investigating potential interventions on disruptive impacts of Industry 4.0 technologies in circular supply chains: Evidence from SMEs of an emerging economy. *Comput. Ind. Eng.* **2022**, *174*, 108753. [CrossRef]
21. Smiatek, J.; Clemens, C.; Herrera, L.M.; Arnold, S.; Knapp, B.; Presser, B.; Jung, A.; Wucherpfennig, T.; Bluhmki, E. Generic and specific recurrent neural network models: Applications for large and small scale biopharmaceutical upstream processes. *Biotechnol. Rep.* **2021**, *31*, e00640. [CrossRef]
22. Karim, M.N.; Rivera, S.L. 992 ACCJTM4 use of recurrent neural networks for bioprocess identification in on-line optimization by micro-genetic algorithms. In Proceedings of the 1992 American Control Conference, Chicago, IL, USA, 24–26 June 1992.
23. Mbiki, S.; McClendon, J.; Alexander-Bryant, A.; Gilmore, J. Classifying changes in LN-18 glial cell morphology: A supervised machine learning approach to analyzing cell microscopy data via FIJI and WEKA. *Med Biol. Eng. Comput.* **2020**, *58*, 1419–1430. [CrossRef] [PubMed]
24. Wang, X.; Zhou, G.; Liang, L.; Liu, Y.; Luo, A.; Wen, Z.; Wang, X.Z. Deep learning-based image analysis for in situ microscopic imaging of cell culture process. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107621. [CrossRef]
25. Guo, W.; Liu, X.; Xiang, L. Membrane System-Based Improved Neural Networks for Time-Series Anomaly Detection. *Processes* **2020**, *8*, 1168. [CrossRef]
26. Hemavathi, B.; Vidya, G.; Anantharaju, K.S.; Pai, R.K. Machine learning in the era of smart automation for renewable energy materials. *e-Prime-Adv. Electr. Eng. Electron. Energy* **2024**, *7*, 100458. [CrossRef]
27. Xiao, Z.; Li, W.; Moon, H.; Roell, G.W.; Chen, Y.; Tang, Y.J. Generative Artificial Intelligence GPT-4 Accelerates Knowledge Mining and Machine Learning for Synthetic Biology. *ACS Synth. Biol.* **2023**, *12*, 2973–2982. [CrossRef] [PubMed]
28. Ozturk, S.S. Opportunities and challenges for the implementation of continuous processing in biomanufacturing. In *Continuous Processing in Pharmaceutical Manufacturing*; Wiley: Hoboken, NJ, USA, 2015; pp. 457–478. [CrossRef]
29. Farid, S.S. Process economics of industrial monoclonal antibody manufacture. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2007**, *848*, 8–18. [CrossRef] [PubMed]
30. Vogel, J.H.; Nguyen, H.; Giovannini, R.; Ignowski, J.; Garger, S.; Salgotra, A.; Tom, J. A new large-scale manufacturing platform for complex biopharmaceuticals. *Biotechnol. Bioeng.* **2012**, *109*, 3049–3058. [CrossRef] [PubMed]
31. Bielser, J.-M.; Wolf, M.; Souquet, J.; Broly, H.; Morbidelli, M. Perfusion mammalian cell culture for recombinant protein manufacturing—A critical review. *Biotechnol. Adv.* **2018**, *36*, 1328–1340. [CrossRef]
32. Kim, H.S.; Lee, G.M. Differences in optimal pH and temperature for cell growth and antibody production between two Chinese hamster ovary clones derived from the same parental clone. *J. Microbiol. Biotechnol.* **2007**, *17*, 712–720.
33. Harcum, S.W.; Elliott, K.S.; Skelton, B.A.; Klaubert, S.R.; Dahodwala, H.; Lee, K.H. PID controls: The forgotten bioprocess parameters. *Discov. Chem. Eng.* **2022**, *2*, 1. [CrossRef]
34. Routledge, S.J. Beyond De-Foaming: The Effects of Antifoams on Bioprocess Productivity. *Comput. Struct. Biotechnol. J.* **2012**, *3*, e201210001. [CrossRef] [PubMed]
35. Pan, X.; Streefland, M.; Dalm, C.; Wijffels, R.H.; Martens, D.E. Selection of chemically defined media for CHO cell fed-batch culture processes. *Cytotechnology* **2016**, *69*, 39–56. [CrossRef] [PubMed]
36. McDonnell, S.; Floyd Principe, R.; Soares Zamprognio, M.; Whelan, J. Challenges and Emerging Technologies in Biomanufacturing of Monoclonal Antibodies (mAbs). 2022. Available online: www.intechopen.com (accessed on 27 February 2024).
37. Jayapal, K.P.; Wlaschin, K.F.; Hu, W.; Yap, M.G. Recombinant Protein Therapeutics from CHO Cells—20 Years and Counting. *Chem. Eng. Prog.* **2007**, *103*, 40.
38. Hossler, P.; Khattak, S.F.; Li, Z.J. Optimal and consistent protein glycosylation in mammalian cell culture. *Glycobiology* **2009**, *19*, 936–949. [CrossRef]
39. Walsh, G. Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* **2018**, *36*, 1136–1145. [CrossRef] [PubMed]
40. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection: A data perspective. *ACM Comput. Surv.* **2017**, *50*, 3136625. [CrossRef]
41. García, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl. -Based Syst.* **2016**, *98*, 1–29. [CrossRef]

42. Çetin, V.; Yıldız, O. A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Univ. J. Eng. Sci.* **2022**, *28*, 299–312. [CrossRef]
43. Anowar, F.; Sadaoui, S.; Selim, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput. Sci. Rev.* **2021**, *40*, 100378. [CrossRef]
44. Li, C. Preprocessing Methods and Pipelines of Data Mining: An Overview. Seminar Data Mining, Jun. 2019. Available online: http://arxiv.org/abs/1906.08510 (accessed on 17 March 2024).
45. Walsh, I.; Myint, M.; Nguyen-Khuong, T.; Ho, Y.S.; Ng, S.K.; Lakshmanan, M. Harnessing the potential of machine learning for advancing "Quality by Design" in biomanufacturing. *mAbs* **2022**, *14*, 2013593. [CrossRef] [PubMed]
46. Bayrak, E.S.; Wang, T.; Cinar, A.; Undey, C. Computational Modeling of Fed-Batch Cell Culture Bioreactor: Hybrid Agent-Based Approach. *IFAC-PapersOnLine* **2015**, *48*, 1252–1257. [CrossRef]
47. Yatipanthalawa, B.S.; Fitzsimons, S.E.W.; Horning, T.; Lee, Y.Y.; Gras, S.L. Development and validation of a hybrid model for prediction of viable cell density, titer and cumulative glucose consumption in a mammalian cell culture system. *Comput. Chem. Eng.* **2024**, *184*, 108648. [CrossRef]
48. Ebersbach, H.; Geisse, S. Minimizing immunogenicity of biopharmaceuticals by controlling critical quality attributes of proteins. *Biotechnol. J.* **2012**, *7*, 1433–1443. [CrossRef]
49. Medlock, G.L.; Papin, J.A. Guiding the Refinement of Biochemical Knowledgebases with Ensembles of Metabolic Networks and Machine Learning. *Cell Syst.* **2020**, *10*, 109–119.e3. [CrossRef]
50. Shek, C.F.; Kotidis, P.; Betenbaugh, M. Mechanistic and data-driven modeling of protein glycosylation. *Curr. Opin. Chem. Eng.* **2021**, *32*, 100690. [CrossRef]
51. Powers, D.N.; Velugula-Yellela, S.R.; Trunfio, N.; Angart, P.; Faustino, A.; Agarabi, C. Automated Microbioreactors and the Characterization of Media Dependent Changes in Antibody Product Glycosylation and Aglycosylation. *J. Glycobiol.* **2018**, *7*, 1000133. [CrossRef]
52. Zürcher, P.; Sokolov, M.; Brühlmann, D.; Ducommun, R.; Stettler, M.; Souquet, J.; Jordan, M.; Broly, H.; Morbidelli, M.; Butté, A. Cell culture process metabolomics together with multivariate data analysis tools opens new routes for bioprocess development and glycosylation prediction. *Biotechnol. Prog.* **2020**, *36*, e3012. [CrossRef]
53. Le, H.; Kabbur, S.; Pollastrini, L.; Sun, Z.; Mills, K.; Johnson, K.; Karypis, G.; Hu, W.-S. Multivariate analysis of cell culture bioprocess data—Lactate consumption as process indicator. *J. Biotechnol.* **2012**, *162*, 210–223. [CrossRef]
54. Tulsyan, A.; Huang, B.; Gopaluni, R.B.; Forbes, J.F. Performance assessment, diagnosis, and optimal selection of non-linear state filters. *J. Process. Control.* **2014**, *24*, 460–478. [CrossRef]
55. Tulsyan, A.; Khare, S.; Huang, B.; Gopaluni, B.; Forbes, F. A switching strategy for adaptive state estimation. *Signal Process.* **2018**, *143*, 371–380. [CrossRef]
56. Bayrak, E.S.; Wang, T.; Tulsyan, A.; Coufal, M.; Undey, C. Product Attribute Forecast: Adaptive Model Selection Using Real-Time Machine Learning. *IFAC-PapersOnLine* **2018**, *51*, 121–125. [CrossRef]
57. Cesmat, J.; McAndrew, J. The Significance of PID Tuning within Biopharmaceutical Processes. *White Pap.* **2022**, *223*, 1–8.
58. Foley, M.W.; Julien, R.H.; Copeland, B.R. A Comparison of PID Controller Tuning Methods. *Can. J. Chem. Eng.* **2005**, *83*, 712–722. [CrossRef]
59. Park, S.-Y.; Choi, D.-H.; Song, J.; Park, U.; Cho, H.; Hong, B.H.; Shozui, F.; Silberberg, Y.R.; Lee, D.-Y. Characterizing Basal and Feed Media Effects on Mammalian Cell Cultures by Systems Engineering Approaches. *IFAC-PapersOnLine* **2022**, *55*, 31–36. [CrossRef]
60. Rashedi, M.; Rafiei, M.; Demers, M.; Khodabandehlou, H.; Wang, T.; Tulsyan, A.; Undey, C.; Garvin, C. Machine learning-based model predictive controller design for cell culture. *Biotechnol. Bioeng.* **2023**, *120*, 2045–2377. [CrossRef] [PubMed]
61. Grilo, A.L.; Mantalaris, A. Apoptosis: A mammalian cell bioprocessing perspective. *Biotechnol. Adv.* **2019**, *37*, 459–475. [CrossRef] [PubMed]
62. Müller, S.; Harms, H.; Bley, T. Origin and analysis of microbial population heterogeneity in bioprocesses. *Curr. Opin. Biotechnol.* **2010**, *21*, 100–113. [CrossRef]
63. Kiesslich, S.; Kamen, A.A. Vero cell upstream bioprocess development for the production of viral vectors and vaccines. *Biotechnol. Adv.* **2020**, *44*, 107608. [CrossRef]
64. Drugmand, J.-C.; Schneider, Y.-J.; Agathos, S.N. Insect cells as factories for biomanufacturing. *Biotechnol. Adv.* **2012**, *30*, 1140–1157. [CrossRef]
65. Xie, D. Continuous biomanufacturing with microbes—Upstream progresses and challenges. *Curr. Opin. Biotechnol.* **2022**, *78*, 102793. [CrossRef] [PubMed]
66. Peternel, Š. Bacterial cell disruption: A crucial step in protein production. *New Biotechnol.* **2013**, *30*, 250–254. [CrossRef] [PubMed]
67. Oyetunde, T.; Liu, D.; Martin, H.G.; Tang, Y.J. Machine learning framework for assessment of microbial factory performance. *PLoS ONE* **2019**, *14*, e0210558. [CrossRef]
68. Malla, R.; Srilatha, M.; Farran, B.; Nagaraju, G.P. mRNA vaccines and their delivery strategies: A journey from infectious diseases to cancer. *Mol. Ther.* **2024**, *32*, 13–31. [CrossRef] [PubMed]
69. Fernandes, P.; Silva, A.C.; Coroadinha, A.S.; Alves, P.M. Upstream bioprocess for adenovirus vectors. In *Adenoviral Vectors for Gene Therapy*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 2016; pp. 139–161. [CrossRef]

70. Collins, L.T.; Ponnazhagan, S.; Curiel, D.T. Synthetic Biology Design as a Paradigm Shift toward Manufacturing Affordable Adeno-Associated Virus Gene Therapies. *ACS Synth. Biol.* **2023**, *12*, 17–26. [CrossRef] [PubMed]
71. Daneshvar, A.; Lukianov, S.N. Artificial Intelligence-Mediated Computer-Aided Design of Viral Gene Therapies. *GEN Biotechnol.* **2023**, *2*, 482–489. [CrossRef]
72. Quintanilla, R.H.; MacArthur, C.C.; Fergus, J.K.; Lakshmipathy, U. Clinical Translation of Vector Production and Protocol Preparation I 219. Use of Novel Surface Markers to Track the Kinetics of Somatic Cell Reprogramming. *Mol. Ther.* **2014**, *22*, S84.
73. Marwidi, Y.; Nguyen, H.-O.B.; Santos, D.; Wangzor, T.; Bhardwaj, S.; Ernie, G.; Prawdzik, G.; Lew, G.; Shivak, D.; Trias, M.; et al. A robust and flexible baculovirus-insect cell system for AAV vector production with improved yield, capsid ratios and potency. *Mol. Ther. Methods Clin. Dev.* **2024**, *32*, 101228. [CrossRef]
74. Altenburg, J.J.; Klaverdijk, M.; Cabosart, D.; Desmecht, L.; Brunekreeft-Terlouw, S.S.; Both, J.; Tegelbeckers, V.I.P.; Willekens, M.L.P.M.; van Oosten, L.; Hick, T.A.H.; et al. Real-time online monitoring of insect cell proliferation and baculovirus infection using digital differential holographic microscopy and machine learning. *Biotechnol. Prog.* **2023**, *39*, e3318. [CrossRef]
75. Angelov, P.P.; Soares, E.A.; Jiang, R.; Arnold, N.I.; Atkinson, P.M. Explainable artificial intelligence: An analytical review. *WIREs Data Min. Knowl. Discov.* **2021**, *11*, e1424. [CrossRef]
76. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Mark*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. [CrossRef]