Author/s:
Karimi Alavijeh, M;Lee, YY;Gras, SL

Title:
A perspective-driven and technical evaluation of machine learning in bioreactor scale-up: A case-study for potential model developments

Date:
2024-07-01

Citation:
Karimi Alavijeh, M., Lee, Y. Y. & Gras, S. L. (2024). A perspective-driven and technical evaluation of machine learning in bioreactor scale-up: A case-study for potential model developments. Engineering in Life Sciences, 24 (7), https://doi.org/10.1002/elsc.202400023.

Persistent Link:
https://hdl.handle.net/11343/351093

Received: 29 February 2024 | Accepted: 1 March 2024

DOI: 10.1002/elsc.202400023

Engineering in Life Sciences

RESEARCH ARTICLE

A perspective-driven and technical evaluation of machine learning in bioreactor scale-up: A case-study for potential model developments

Masih Karimi Alavijeh[1,2] | **Yih Yean Lee**[3] | **Sally L. Gras**[1,2]

[1]Department of Chemical Engineering, The University of Melbourne, Parkville, Victoria, Australia

[2]The Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, Victoria, Australia

[3]CSL Innovation, Melbourne, Victoria, Australia

Correspondence
Sally L. Gras, Department of Chemical Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.
Email: sgras@unimelb.edu.au

Abstract

Bioreactor scale-up and scale-down have always been a topical issue for the biopharmaceutical industry and despite considerable effort, the identification of a fail-safe strategy for bioprocess development across scales remains a challenge. With the ubiquitous growth of digital transformation technologies, new scaling methods based on computer models may enable more effective scaling. This study aimed to evaluate the potential application of machine learning (ML) algorithms for bioreactor scale-up, with a specific focus on the prediction of scaling parameters. Factors critical to the development of such models were identified and data for bioreactor scale-up studies involving CHO cell-generated mAb products collated from the literature and public sources for the development of unsupervised and supervised ML models. Comparison of bioreactor performance across scales identified similarities between the different processes and primary differences between small- and large-scale bioreactors. A series of three case studies were developed to assess the relationship between cell growth and scale-sensitive bioreactor features. An embedding layer improved the capability of artificial neural network models to predict cell growth at a large-scale, as this approach captured similarities between the processes. Further models constructed to predict scaling parameters demonstrated how ML models may be applied to assist the scaling process. The development of data sets that include more characterization data with greater variability under different gassing and agitation regimes will also assist the future development of ML tools for bioreactor scaling.

KEYWORDS
bioprocessing, bioreactor, data-driven modeling, machine learning, mammalian cell

Abbreviations: ANN, artificial neural networks; CFD, computational fluid dynamics; CHO, Chinese hamster ovary; Dt, vessel diameter; Di, impeller diameter; IVCD, integral viable cell density; IQR, interquartile range; kLa, gas mass transfer coefficient; mAb, monoclonal antibody; ML, machine learning; PCA, principal component analysis; PC, principal component; P/V, power input per volume; Re, Reynolds number; RMSE, root-mean-square error; SHAP, SHapley Additive exPlanations; VCD, viable cell density; VIF, variance inflation factor; XGBoost, eXtreme Gradient Boosting.

1 | INTRODUCTION

The global market for therapeutic monoclonal antibody (mAb) products was ~$163 billion in 2019, having grown faster than all other biopharmaceutical products in recent years; matched by an increase in global manufacturing capacity from 10 to 25 tonnes between 2013 and 2019 [1]. The ability to scale production of new mAb products is critical to this expansion and requires novel scale-up solutions, since traditional techniques are not always accurate, reproducible or transferrable across processes.

The comparison of bioreactors across scales is complex, with biological phenomena linked to a plethora of interactions among multiple variables. These include engineering scale-dependent variables, scale-independent variables, and cell-related variables, along with unknown reaction kinetics and nonlinear and dynamic behaviors that feature time-varying parameters [2].

Engineering factors and the design configurations of bioreactors inevitably vary from small scale to production scale. These scale-dependent factors can significantly influence flow patterns and mixing regimes within a bioreactor resulting in different:

(i) local distribution profiles (e.g., nutrients, metabolites, pH, temperature, shear, flow velocity, and energy dissipation rates) and the degree of homogeneity;
(ii) mass transfer rates and gas solubilities, in particular O_2 transfer rate and CO_2 stripping rate; and
(iii) hydrodynamic shear-induced cell damage and bubble-induced cell damage.

All of which makes predicting cell behavior as a function of bioreactor scale difficult.

Despite much research developing systematic approaches to facilitate bioreactor scale-up [3–6], no universal solution has been found to overcome the process challenges faced in the development and scale-up of new mAb production processes. It is not possible to maintain all scaling parameters at constant values between scales. Methods based on only one scaling parameter, such as equal power densities, equal oxygen volumetric mass transfer rates, or equal tip speeds between two scales, are also at a risk of failure, as other parameters will change with scale.

A digital approach has the potential to increase our understanding of the scaling process and not only aid process development by scaling experts but also to assist scientists with different expertise to identify suitable scaling parameters, de-risking process scaling and troubleshooting. A number of studies have examined how computer-based methodologies can assist bioreactor scale-up, as reviewed recently [2]. Although the techniques vary widely in terms of their practical application, they can be classified into three main categories: mechanistic modeling, data-driven modeling, and a third category of hybrid modeling, which combines these former two approaches.

Mechanistic models describing mammalian cell culture within a bioreactor have been developed in a number studies [7–11]. These can provide useful insights; however, their specific application to predict scaling parameters for the purposes of scaling is still limited, due to the lack of mechanistic equations describing the effect of scale-dependent variables on cell behavior. Moreover, limited knowledge of the underlying mechanisms, overparameterization, and the uncertainty associated with parameter estimation, make this knowledge-based approach restricted when applied across scales.

Computational Fluid Dynamics (CFD) can also contribute to mechanistic knowledge of cell culture conditions to assist in scaling and this approach allows scaling factors to be integrated into the knowledge-based analysis of cell culture at different bioreactor scales. Several CFD studies have already predicted a set of suitable scaling parameters through detailed evaluation of the flow fields and bioreactor hydrodynamics, ensuring equivalent performance between scales for a specific process [12–14]. These models focus on one process with a particular bioreactor geometry, although similar governing equations can be used to model multiphase flows in different scale-up processes with differing geometries. A key drawback of this approach has been the significant knowledge and expertise associated with the choice of CFD models, as well as the computational cost involved. Nevertheless, the emergence of Graphics Processing Unit (GPU) accelerated CFD tools that can run on a single, reasonably priced desktop GPU will enhance accessibility by significantly increasing computational performance while lowering costs. Additional work is also required to build transferable models that can be used by scientists with different backgrounds.

A further advance has been the development of novel compartmental modeling frameworks that combine cell kinetics with CFD in order to incorporate local hydrodynamic characteristics derived from CFD simulations into biokinetic models, rather than assuming an ideally uniform environment for cellular growth within an agitated vessel [15]. This approach is based on generating several zones over the entire volume of the bioreactor, with each zone being treated as a hypothetical well-mixed bioreactor in which the kinetic model is solved. Although this approach has shown promising results [16], it is acknowledged that experimental measurements of local cell populations for validation purposes is difficult. It is anticipated, however, that advances in experimental and modeling techniques will make compartmental models more applicable to the in-depth evaluation of cell growth

across scales, in particular at large scales where culture heterogeneities may be more significant.

The methods discussed above, particularly CFD models, have been successful in addressing the problem scope, that is, identifying suitable scaling parameters that can be applied to assist scaling. The practical evaluation of data-driven models that can establish multiparameter scaling tools, however, is limited. Machine learning (ML) can be considered as a potential complementary approach to reduce this gap. With the aid of ML models, knowledge captured in bioprocess data can potentially be transferred from a smaller scale bioreactor to a larger scale bioreactor. There is also potential to transfer knowledge across one different process to another. A key advantage of this approach is the construction of mathematical relationships between scaling parameters and key performance indicators without the need for the inclusion of underlying mechanisms, most of which are not well understood. ML algorithms can also provide scientists with deeper insights into the interactions among key variables and potential sources of variation to minimize the risk of failure during bioreactor scale-up.

Despite the promise of data-based approaches, the application of such methods for the prediction of process performance and scaling factors is still in the early stages of development. Research is therefore needed to understand the potential of ML models to predict both process performance and scaling parameters, such as for mAb production in CHO cells, to identify potential opportunities and limitations, as well as future research challenges. This study therefore sought to collate the available data for these production processes and to assess the potential of ML techniques to predict both process performance and scaling parameters.

2 | METHODOLOGY

The overall methodology applied in this study includes the identification of factors important to scaling, a literature search, data collection and extraction, data preprocessing, development of supervised and unsupervised models, and interpretability analysis that are described in detail in the following sections.

2.1 | Augmented feature selection by domain knowledge

Knowledge from the bioreactor scale-up domain was incorporated into the feature selection process. This involved analyzing the key variables affecting bioreactor scale-up and identifying relevant features that are technically significant, actionable, and available through experiments or calculations, as well as being physically meaningful for real-world industrial application. The methodology included a comprehensive literature review and refinement of the feature space.

2.2 | Data collection procedure

A thorough study of the published literature was performed to collect bioreactor scale-up data for mammalian CHO cells expressing antibody products. To this end, different combinations of keywords were used to find relevant public sources in Google, Scopus, Web of Science, as well as reports provided by the biopharmaceutical industry, including Cytiva, Eppendorf, Merck Millipore, Thermo Fisher Scientific, Sartorius, Allegro, and Applikon. Relevant reports characterizing bioreactor performance [17–33] were also used to obtain the details of bioreactor designs where needed.

A total of 18 different processes, comprising 755 time-series measurements obtained from 55 bioreactors at various scales from 250 mL to 5000 L were compiled into the dataset. The scales included in each process are shown in Table 1S in the Supplementary Information. Both scale-dependent factors—including process scale (volume), aspect ratio, impeller diameter, vessel diameter, power input, impeller tip speed, volumetric oxygen mass transfer, Kolmogorov length scale, Reynolds number, mixing time, agitation rate were included. Scale-independent variables included pH setpoint, temperature setpoint, dissolved oxygen setpoint, seeding (or inoculation) density, and culture duration. The dataset also contained cell growth-related parameters, including cell viability, viable cell density (VCD), lactate concentration, specific growth rate, and integral viable cell density (IVCD).

Most VCD, viability and lactate concentration data were represented in figures rather than being tabulated; thus, WebPlotDigitizer, an online data extraction tool (https://apps.automeris.io/wpd/), was used to collect data from the figures where needed. The specific growth rates were calculated during the exponential growth phase of the cultures using the data of viable cell density as a function of time. IVCD is an important metric in cell culture that is directly related to the specific rates of product formation, nutrient consumption and cell growth. The cumulative viable cell concentration over the culture period up to time t—denoted by $IVCD_t$—was calculated using the trapezoidal rule by Equation (1): [34]

$$IVCD_t = IVCD_{t-1} + \frac{VCD_t + VCD_{t-1}}{2} \times \Delta t \quad (1)$$

2.3 | Data pre-processing

Box plots were used to visualize the statistical distribution of each variable in the raw dataset and to identify possible outliers. No outlier was detected for the target variable of VCD. Statistically identified outliers for independent variables that are scaling parameters, such as agitation rate, power input, and bioreactor aspect ratio were included in the dataset, as their variation is commonly expected across scales and is expected to be physically meaningful based on prior knowledge and technical reports [28].

Data imputation was not performed, as the data obtained were from different processes with a limited number of experiments, or observations. Features that were reported across all processes were selected for inclusion in the preprocessed database. Conversely, features where data were not reported across all processes (e.g., >50% of oxygen mass transfer coefficients were not reported across all processes) were excluded from the dataset. There were no missing data within each process for the selected features. Since not all the observations belong to the same process, a process name (e.g., Process 1, Process 2, etc.) was included as a categorical variable in the dataset to account for different process characteristics, including cell line, media, or feeding strategy. This categorical variable was converted to a numerical input for ML algorithms using one-hot encoding, where a value equal to one was assigned for observations within the same process, while zero was assigned to other processes.

A standardization method was then applied to ensure that features with greater values would not dominate over those with smaller values. Equation (2) was used to transform each variable X based on the mean ($\bar{X}$) and standard deviation (σ) [35, 36]. This transformation was performed by the StandardScaler method in the Scikitlearn library in Python.

$$Z = \frac{X - \bar{X}}{\sigma} \tag{2}$$

2.4 | Comparing processes using principal component analysis

Principal component analysis (PCA) was used as a method of unsupervised multivariate data analysis to assess and potentially explain variance in the original dataset. This transformation of data into a low-dimensional space can facilitate statistical comparisons between processes and was used to assess similarity between the bioprocesses by identifying the location of clusters in the bioreactor data on the score plot of the first two principal components that explain the highest variability. In this case, the variables of bioreactor scale, including scale (volume), vessel diameter (D_t), aspect ratio, impeller diameter (D_i), D_i/D_t, power input per volume (P/V), impeller tip speed, Kolmogorov length, Reynolds number (Re), mixing time, agitation rate, seeding density, culture duration at peak VCD, final viability, peak VCD, final IVCD, and specific growth rate were examined.

The evolution of a bioreactor culture with time can also be studied using PCA, as described previously [37, 38]. A second PCA was therefore performed to detect any significant change in cell metabolism during cultivation within each process by assessing the change in cell density, viability, IVCD, and lactate concentrations as a function of time. The FactoMineR package [39] and pca3d package [40] in R were used to conduct the PCA, applying a confidence ellipse level of 95%.

2.5 | Machine learning regression model development

The complexity of predicting cell behavior within a bioreactor at different scales is not only attributed to intracellular phenomena but also linked to changes in scale-sensitive variables that act as external factors influencing cell growth and metabolism. This includes the impact of scale-sensitive parameters on cell signaling pathways, for which there are limited mechanistic mathematical descriptions [11, 41]. Traditionally, to mitigate this complexity, a suitable scale-up criterion (commonly P/V) is maintained constant between scales, the equivalence between scales is then examined by comparing the plots of daily measurements (such as daily VCD and titre) from experiments across scales. For a successful bioreactor scale-up, ideally VCD, or other critical measurements, should be equal across the two scales on each day (t) as follows:

$$(VCD_t)_{Scale1} = (VCD_t)_{Scale2} \tag{3}$$

This process is always at the risk of failure, as the importance of other scaling parameters is not completely captured in this strategy. For example, the surface (or overlay) gas exchange and bubble gas exchange for CO_2 and O_2 have a substantial impact on cellular growth in a bioreactor. Although P/V is correlated with gas transfer, an equivalent P/V does not necessarily provide equivalent gas transfer; it has been shown that surface exchange is correlated with Reynolds number and aspect ratio, while bubble gas transfers are mainly correlated with P/V and gas flowrate [42]. Furthermore, such interdependencies are not the same between two different scales.

An alternative strategy is to search for a set of values for multi scaling parameters in an acceptable range that

when applied together ensure equivalent performance. Nonetheless, owing to a lack of mechanistic information, as explained above, achieving a set of such values is not a trivial task. With this in mind, we set out to examine the potential of ML predictive models to describe the relationship between scaling parameters and a target measurement, such as VCD, as shown by Equation (4):

$$VCD = f\left(t, \frac{P}{V}, \textit{aspect ratio}, \textit{Reynolds number}, \textit{impeller tip speed}, \textit{mixing time}, \ldots\right) \quad (4)$$

In this study, we employed and compared data-driven models, namely artificial neural networks and a gradient boosting method, to help build mathematical relationships between such scale-sensitive variables.

The full dataset of 18 processes with 755 datapoints collected from 55 bioreactors, preprocessed as described in Section 2.2, was used for machine learning. The features used as model inputs were ten scale sensitive variables: scale (volume), vessel diameter (D_t), aspect ratio, impeller diameter (D_i), D_i/D_t, power input per volume (P/V), impeller tip speed, Reynolds number (Re), mixing time, agitation rate, and a further feature of bioreactor seeding density. VCD was selected as the response variable (the model output), as this was a key output reported in all the public sources reviewed. Other important performance indicators, including product titre, key metabolite concentrations, and product critical quality attributes can also be used as response variables in a similar way but in this instance, most of these outputs were not available in the publicly available data on bioreactor scale-up.

2.5.1 | Model development: artificial neural network

Artificial neural networks (ANNs) have widely been used to model complicated biological processes [43–46] due to their great flexibility in modeling interconnectivities between independent and response variables. An ANN model was therefore developed and the ability of this model to predict VCD across scales between different processes assessed. As there was no prior knowledge or rules available for the optimum ANN topology, a hyperparameter optimization procedure was established to construct the ANN model. A three-layer ANN comprised of an input layer, one hidden layer and an output layer was developed in the TensorFlow Keras API in Python 3.10. The network hyperparameters included the number of neurons in the hidden layer, activation function, learning algorithm, learning rate, batch size, and number of epochs. In addition to this conventional ANN, the application of embedding neural networks that are extensively utilized in Natural Language Processing (NLP) to capture contextual and semantic relationships between words, was examined to transfer knowledge between scaling processes.

A Bayesian optimization technique was implemented to tune the hyperparameters using a 5-fold cross-validation repeated five times (performed by the repeated K-Fold cross validator in the Scikitlearn library) [47, 48]. In this procedure, the data were randomly split into five data subsets and the model then trained on four of the five subsets. The network weights and biases were obtained by minimizing the mean squared error (or loss function) between the experimental and modeled VCDs. Afterwards, the unused subset of data was utilized as a test set to evaluate the performance of the model in each iteration of the cross-validation procedure. After 25 iterations with random combinations of training and test subsets for each set of hyperparameters, the average loss function was calculated. The best hyperparameters were then identified, where the lowest value of the loss function was obtained. An EarlyStopping callback was also considered to monitor the loss function and to avoid overfitting once no further improvement in the loss function minimization was achieved. The optimized structure of the ANN was then used to train the model on the standardized bioreactor scaling dataset. In all training procedures, the dataset was split with 70% allocated for the training set and 30% allocated for the validation set.

As described above, an ANN topology with one additional embedding layer was also developed to capture potential similarity among the bioreactor processes examined. A one-hot encoded vector was not sufficiently effective, as it is sparse (i.e., most cells are zero); further, it cannot convey any information regarding the similarity between different processes. An embedding layer, on the other hand, automatically generates a D-dimensional vector of numerical values representing each process; hence, capturing the possible relationships among different bioreactor scale-up processes. This procedure has recently been employed to successfully transfer knowledge across different cell lines [49]. The dimension of the embedding vector (D) was assigned as a new hyperparameter and was determined using the same optimization method applied to determine other network hyperparameters, as described above.

2.5.2 | Model development: XGBoost (eXtreme Gradient Boosting)

With its ability to model nonlinear problems, XGBoost has been shown to be an effective ensemble machine learning

algorithm that can be applied to biological systems [50–55]. This particular algorithm involves parallel trees to provide greater predictive performance and scalability, with lower computational requirements compared to neural networks, it also capable of providing feature importance [56]. Like the ANN model development explained above, a Bayesian hyperparameter tuning procedure with 5-fold cross-validation was developed to create XGBoost models using the XGBoost Python package. The hyperparameters included the step size shrinkage (eta), minimum loss reduction (gamma), learning rate, subsample ratio, number of boosting iterations, and the maximum depth of a tree.

2.5.3 | ML case studies

The predictive performance of ML algorithms described above was initially evaluated using test data selected randomly from different scale-up processes from the full preprocessed dataset. This scenario is somewhat different to an actual scale-up problem, where a specific process, that is, individual cell line and product, is scaled for larger scale production and the aim is to examine key performance indicators across scales but within the same process and product. Additional case studies were therefore considered to specifically evaluate the performance of the ML models during bioreactor scaling, including for the same cell line, process, and product (Table 1).

In the first case study, the ability of the model to predict across processes was tested. All data for a selected process of interest were removed from the dataset (e.g., Process 1 or Process 2, etc. where only one process was removed at a time). To prevent leakage, a new ML algorithm was then optimized and trained on the remaining dataset without the removed process. The predictive performance of the model was next tested using the removed data as an unseen new process (see Table 1 for a description of training and test datasets). This approach assumes that the input features are reasonably process independent.

In the second and third case studies, a different approach was taken to better represent scale up in industry and test the ability of the ML algorithms to predict across scale. Typically, most data are initially obtained in small-scale bioreactors, where data acquisition is more feasible and cheaper. The ability to scale based on this small-scale data is therefore important. In these new case studies, all the larger scale data for a specific process were removed from the test dataset during the test phase leaving only the small-scale data for that process. The ML algorithms were then developed using this new small-scale dataset and their ability to predict unseen larger scale data for the same

TABLE 1 Case studies used to test ML algorithm applicability to bioreactor scaling.

Case study	Key question	Training data	Test data	Additional inputs	Model applied and descriptor
#1	**Can the model predict a new unseen process?**	All processes in the dataset except the process of interest for prediction	All data relevant to the process of interest	–	ANN or XGBoost
#2	**Can the model predict across scales in a given process when provided only with small scale data?**	All processes in the dataset plus small-scale data for the process of interest for prediction	All data relevant to the process of interest except for the small-scale data for that process	–	ANN or XGBoost + small scale data
#3	**Can model #2 be improved using entity embeddings (ANN) and one-hot encoding (XGBoost)?**	All processes in the dataset plus small-scale data for the process of interest for prediction	All data relevant to the process of interest except for the small-scale data for that process	Process type is added using entity embeddings (ANN) and one-hot encoding (XGBoost).	ANN + embedding layer or XGBoost (one-hot encoding)

process evaluated (Table 1). This approach again assumed input features are reasonably process independent.

In addition, case study three examined the advantages of adding an embedding layer to the ANN and one-hot encoding to the XGBoost methods of predictions (Table 1).

Model performance was evaluated using the root-mean-square error (RMSE) between the predicted and actual measurements, as follows where N is the number of measurements:

$$RMSE = \sqrt{\frac{\sum (actual\ VCD - predicted\ VCD)^2}{N}} \quad (5)$$

2.5.4 | Machine learning model development for P/V prediction

We also aimed to find a mathematical relationship between P/V and bioreactor geometry (e.g., aspect ratio and vessel diameter), where a target peak viable cell density (VCD) could be obtained from a specific starting seed density, that is:

$$\mathrm{P/V} = \mathrm{f}(\text{seeding density, peak VCD, aspect ratio, vessel diameter, ...}) \quad (6)$$

A total of 12 features, namely scale (volume), vessel diameter (Dt), aspect ratio (H/Dt), impeller diameter (Di), Di/Dt, peak VCD, seeding density, culture duration at peak VCD, impeller tip speed, Reynolds number (Re), mixing time, and agitation rate, could potentially be included in such a model. Since the number of observations (or bioreactor runs) in the dataset is limited to 55, a reduction in dimensionality is required to find the most suitable features to avoid overfitting. To this end, a combination of (1) a distance correlation technique and (2) variance inflation factor (VIF) analysis was used to reduce the number of features. The distance correlation is capable of measuring nonlinear dependencies between two paired variables, so is more suited for the analysis of highly nonlinear systems, such as bioreactors than the Pearson correlation method, which only determines pairwise linear associations between variables [57]. VIF analysis was also used to identify multicollinearity existing among variables, where a VIF of 1 means no multicollinearity [58].

2.5.5 | Determination of feature importance

We calculated SHapley Additive exPlanations (SHAP) values based on the algorithm developed by Lundberg and Lee in 2017 [59]. The SHAP method is a powerful approach that unifies six existing methodologies in the research field of interpretable techniques, which is also reported to be intuitive [59]. The SHAP values were calculated based on the concept of cooperative game theory, in which a weighted average contribution of each feature (or player) to the model output (or payout) is obtained over all possible combinations (or coalitions) of feature values, providing sample-by-sample feature importance and offering local interpretability for each single model prediction [60, 61].

For "F" features, or a maximum coalition size of F, the SHAP explainer model g representing the original model f is defined as a linear model of the feature attribution (or Shapley values) for a feature j (ϕ_j) as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{F} \phi_j z'_j \quad (7)$$

where z' is the coalition binary vector $\in$ {0,1, meaning that when the feature j is present, the corresponding z'_j is 1, otherwise $z'_j = 0$. The goal is to evaluate how model predictions change in the presence or absence of a given feature by estimating the coefficients of the explainer model, that is ϕ'_j. This is estimation is achieved by minimizing the loss function L (Equations (8)):

$$L(f, g, \pi_x) = \sum_{z'} \left[f\left(h_x\left(z'\right)\right) - g\left(z'\right) \right]^2 \pi_x(z') \quad (8)$$

where π_x is the Shapley kernel weight and is calculated by the following equation:

$$\pi_x\left(z'\right) = \frac{F-1}{\binom{F}{|z'|}\,|z'|\,(F-|z'|)} \quad (9)$$

The best explainer model that is very close to f, is then obtained i.e:

$$g\left(z'\right) \approx f\left(h_x\left(z'\right)\right) \quad (10)$$

where h_x is a function that maps the z' vector to corresponding values in the original feature space.

3 | RESULTS AND DISCUSSION

3.1 | The identification of factors important to scaling and data collection

This study started with a review of the literature to identify factors important to scaling CHO mAb production processes. Substantial amounts of time, effort, and cost are commonly dedicated to the effective scaling of bioreactors.

TABLE 2 Important factors that impact cellular behavior within a bioreactor for which limited mechanistic mathematical descriptions are available.

Gas flow rates (superficial gas velocities) including sparger rate and overlay rate
Gas mass transfer coefficients
Sparger design, type, and location
Number of impellers
Impeller design, type, and location
Fluid forces and properties
Probe locations
Agitation rate
Power input
Impeller tip speed
Number of baffles
Vessel baffling configuration
Vessel diameter
Impeller diameter
Gas holdup
Bubble diameters
Eddy sizes
Gas bubble residence time
Kinetic energy dissipation rates
CO_2 stripping time
Mixing time
Shear rates
Other bioreactor design characteristics, such as top and bottom clearance designs.

This is due to many factors that have a direct or indirect impact on cell behavior and the rates of metabolite formation in bioreactors across scales, of which we have little or no mechanistic knowledge. These include but are not limited to the factors listed in Table 2.

In addition to the factors shown in Table 2, process variables such as pH, temperature, dissolved oxygen and carbon dioxide, media composition, seeding density and culture duration can have an impact on key performance indicators, including cell viability, viable cell density, titre, critical product quality attributes, metabolite concentrations, specific growth rate, and specific productivity. Figure 1 presents an Ishikawa diagram where five key upstream categories are presented that can lead to nonequivalent performance between bioreactors at different scales. These are bioreactor design and scale sensitive factors, bioreactor design and hydrodynamics, mass transfer factors, feeding strategy and bioreactor operational conditions, including factors listed in Table 2. In addition, Figure 1 highlights the interconnectivities between these parameters, determined from a review of the literature [41, 62]. For example, the relationship between bioreactor aspect ratio and oxygen mass transfer is complex, making troubleshooting and de-risking highly process-specific and challenging.

Although each factor described above can have an independent influence on a specific performance indicator, it is almost impossible to include all these factors as features in machine learning algorithms, because of the curse of dimensionality (i.e., there are too many features). Furthermore, many of these factors are not typically measured or reported in the literature or may be qualitative, such as the type of impeller/sparger. A comprehensive literature review was therefore conducted to identify factors most crucial to scale up for possible inclusion in a machine learning algorithm:

(i) Parameters including power input per volume, gas mass transfer coefficient (k_La), gas flow rates, impeller tip speed, mixing time, and Reynolds number are often analyzed to obtain insights for matching key performance indicators across bioreactor scales [63–65]. These scale-specific parameters can represent many qualitative and quantitative variables pertaining to mixing, mass transfer, and shear damage to cells.

(ii) There are many factors affecting power consumption in a stirred-tank bioreactor including impeller design and configuration, sparger design and location, fluid properties, and vessel baffling. A power number is then incorporated into power consumption calculations to account for these factors [66–68].

(iii) The Reynolds number characterizing the bioreactor fluid regime represents the ratio of inertial to viscous forces in the culture media, as well as the fluid properties [64, 69] and provides a good characterization of the fluid environment in the bioreactor.

(iv) The scale or production rate is indicated by the bioreactor volume; however, the bioreactor volume is not the sole size-related factor determining flow-field-relevant effects in a bioreactor; various aspect ratios and vessel diameters can be employed for the same vessel volume. The aspect ratio affects gas bubble formation and residence time, leading to changes in the oxygen transfer rate and CO_2 stripping rate [28]. An aspect ratio close to 1 typically results in the improved dissolution of oxygen from air, while aspect ratios up to 3 increase the gas residence time [70]. Aspect ratio also impacts the surface to volume ratio and the hydrostatic pressure, which can influence gas exchange and impact on gas solubility. In addition, it has been shown that although mixing time increases upon scale up, mixing time is also impacted by the aspect ratio at the same volume [70–72].

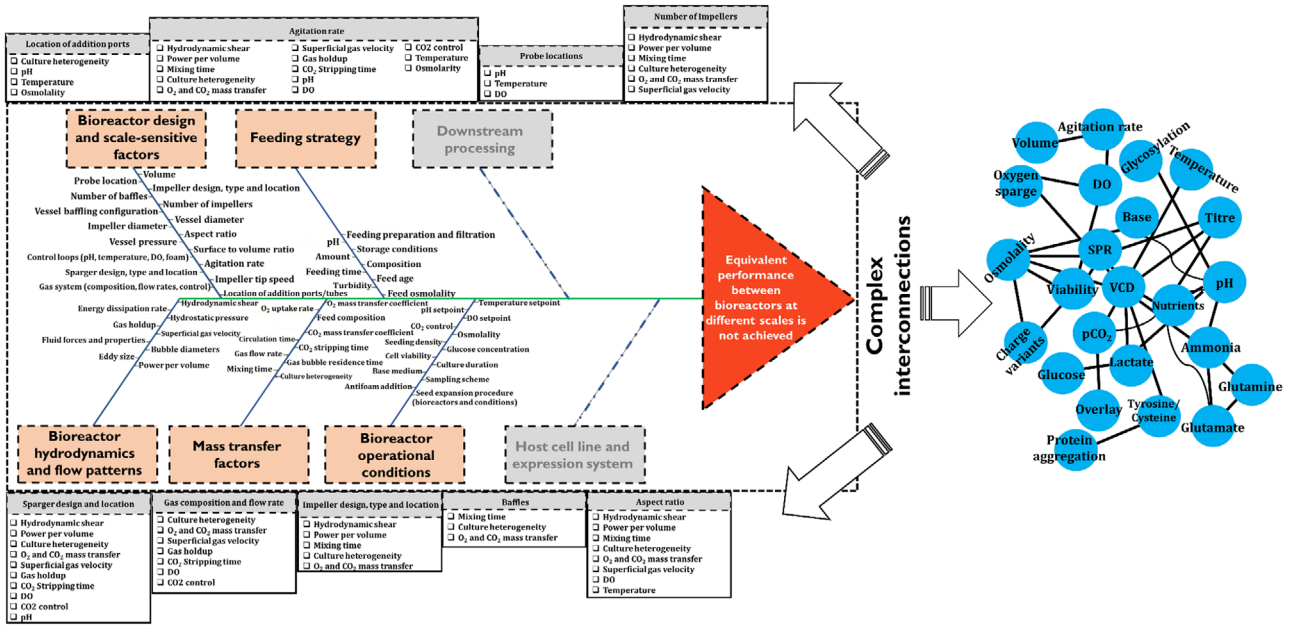

FIGURE 1 An Ishikawa diagram showing possible causes that can contribute to differences in bioreactor performance across scales. These are grouped into five related upstream categories (orange). The complex interconnections among these variables are also listed in tables that appear at the top and bottom (grey) and in the network (adopted from [62]) on the right, which represents the possible impacts and interconnectivities of variables on cell growth and product quality attributes.

(v) Several standard design considerations and special engineering ratios, such as impeller positioning and size, as well as tank baffling and clearance, affect fluid swirling and vortexing and are represented by the vessel diameter [72–74].

(vi) The impeller diameter and its ratio to the vessel diameter are widely used in many characterized formulas employed in bioreactor design and scale-up, in particular for gas-handling capacity estimations, including flooding or loading transitions and gas dispersions [75, 76].

(vii) Many hydrodynamic parameters, such as hydrodynamic shear and gas holdup volume, mass transfer and culture heterogeneity, as well as process variables, such as pH, dissolved O_2 and CO_2 and temperature can vary with agitation rate [77].

Although several scale-dependent parameters are correlated, they can have disparate impacts on cellular behaviors across scales. As a result, relative changes in cell death, growth, metabolism, morphology, and differentiation will depend on the cell type and culture characteristics, requiring customization of process scaling [78, 79]. As a case in point, the gas flow rate has an impact on dissolved oxygen concentrations, because it is directly linked to the bubble residence time. It can also be correlated with power consumption and mass transfer coefficients, as well as CO_2 removal rate [4, 80, 81] CO_2 removal, however, is not affected by the power input [79]. In another example, cell growth can vary between two bioreactors with the same volume but different aspect ratios. The critical factors identified above were therefore all considered in the development of the ML algorithms in this study without performing further dimensionality reduction, unless otherwise stated.

3.2 | Explanatory data analysis

Boxplots were used to explore the data collected for 18 CHO mAb producing processes across 250 mL to 5000 L scale from the literature (Figure 2). These plots show the range of setpoint, dependent, and calculated parameters and provide insightful estimations for the variation in scale dependent process variables.

As controlled variables, the setpoint data for the operational pH and temperature have a low spread (Figure 2A). The seeding density is also easily controlled by operators and is typically held constant across scales, so is quite similar across the dataset for CHO cells. The dissolved oxygen is also typically controlled at 40% across scales in mAb production processes, although it can vary between 30% and 60%.

Significant differences in cellular behavior stemming from internal cellular metabolism and the response to external environmental factors, other than pH and temperature, can be observed in the variation in cell growth related parameters, namely the specific growth rate, final

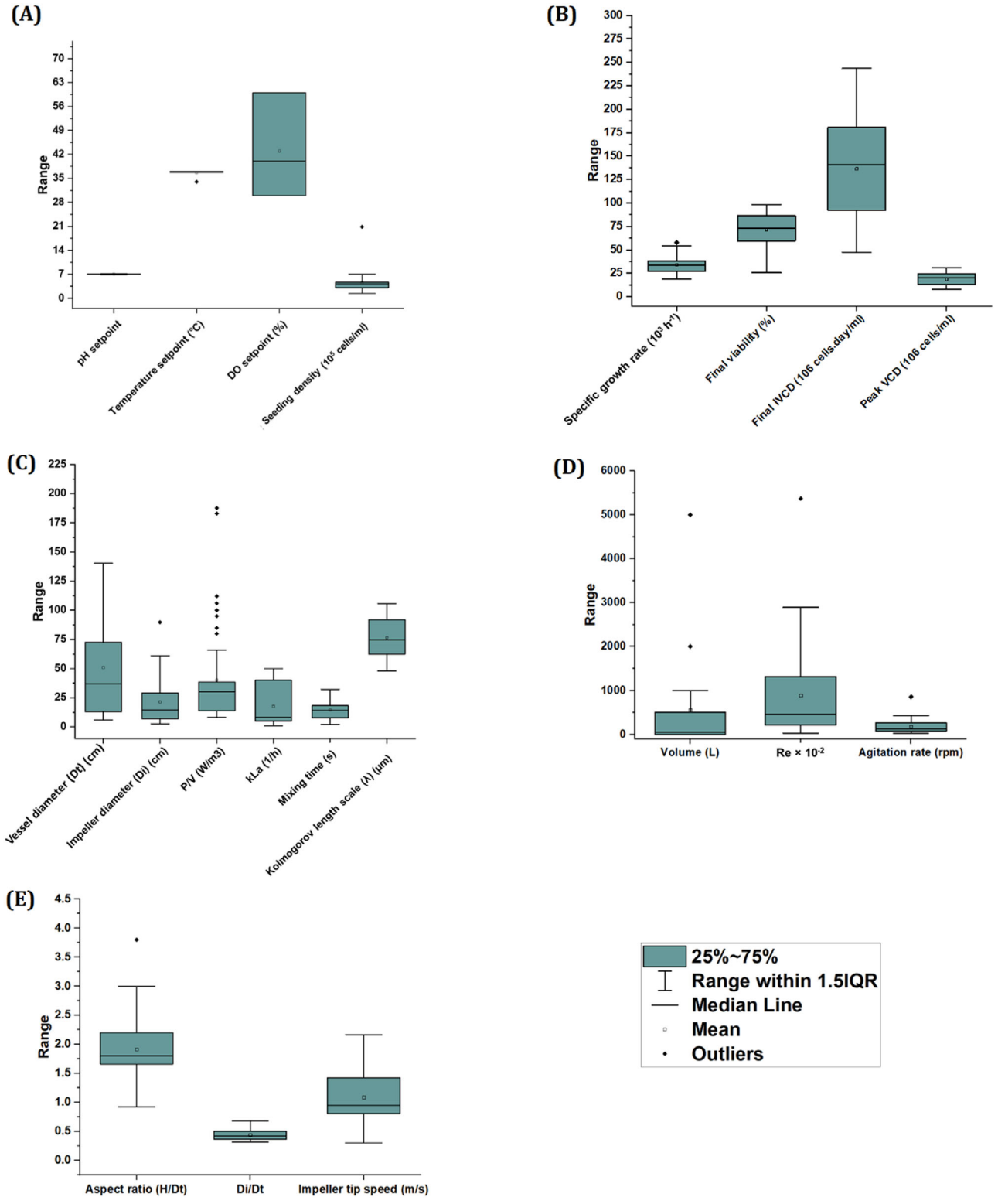

FIGURE 2 Boxplots (A–E) showing the range of key parameters collected for the 18 processes at different scale from literature. Similar parameters with equivalent range are grouped together. IQR in the legend denotes the interquartile range.

viability, final IVCD, and peak VCD. Of these the final IVCD had the greatest range, followed by final viability, both factors that impact on cell and process performance. In contrast, operators often target a similar specific growth rate across experiments and scales. The peak VCD was also similar (Figure 2B). The variations observed are expected, as dynamic cellular behaviors can vary across CHO cell lines in each of the different processes within the dataset.

Several parameters reflect the different scales (250 mL to 5000 L) within the dataset, including the vessel volume and diameter (Figure 2C,D). Other parameters reflect variations in bioreactor design, such as impeller diameter, aspect ratio and D_i/D_t (Figure 2E), although these can also vary with scale. Key process parameters such as agitation rate, impeller tip speed, and mixing time vary but are limited by the physical requirements of the cells.

Several data points are identified as statistical outliers between the different processes (Figure 2); nonetheless, these are not erroneous measurements (erratic outliers) but rather meaningful variations typically expected in bioreactor scaling processes with different physical geometries and operational conditions. The greatest number of outliers was observed for P/V. Power density is an indicator of shear stress and the degree of mixing and homogenization within a bioreactor. Given differences between processes and scales stemming from variations in bioreactor geometries and cell line characteristics, such variations in P/V are expected.

Insights can be obtained from the Reynolds number and the Kolmogorov length scale, or eddy size, which characterize fluid flow within the bioreactor (Figure 2D,E). The mean value of the Reynolds number, which was greater than 80,000, indicates that fully turbulent conditions are commonly established across the mammalian cell bioreactors examined here. Despite turbulent conditions, the Kolmogorov length scale indicates CHO cells are not likely to be damaged. When the average eddy size is greater than the cell size, the cell disruption is not generally expected due to turbulent shear stress [82, 83]. The mean length was 78 μm here, which is much larger than the average diameter of <20 μm for CHO cells, indicating that cell damage due to turbulence would not be a significant issue for the processes studied here. Although the number of processes that reported k_La was limited, an average k_La value of 17.5 h^{-1} was calculated providing insight into these processes.

3.3 | Comparing processes

Principal component analysis (PCA) was performed in order to assess the similarity of the 18 processes and 17 variables (described in Section 2.3), as shown in Figure 3. A plot of the first two uncorrelated principal components, explaining 31.2% and 17.1% of the variance is presented in Figure 3A. Processes located near the center point of the plot, such as process 3, have the most similarity to other processes. Conversely, processes far from the center, such as process 1, are less similar to other processes.

The PCA plot provides insights into processes conducted within Ambr250 bioreactors, an example of a fully automated high-throughput bioreactor system, which represents the smallest scale (250 mL) within the dataset. These efficient systems have frequently been utilized by the pharmaceutical industry for high-throughput process development, particularly for expensive experiments [84, 85] and can improve process understanding and identify optimal conditions [86]. Yet, one consideration often discussed in the literature that is also of interest to companies, is whether such mini bioreactors can be used as qualified scale-down models.

The Ambr250 bioreactor may not be able to perfectly represent the conditions within other larger commercial scale bioreactors, as the data for the four processes using Ambr250 in this dataset are located far from the center, on the left of the PC1 axes, near or out of the confidence ellipse border. This observation is consistent with previous reports that extensive characterization studies are often required to achieve a consistent scale-down model in the Ambr250 system owing to scale-specific limitations that especially affect mass transfer and gassing [87, 88]. Such high-throughput systems can therefore be used as complementary screening platforms but may not effectively represent performance at other scales and advanced risk analysis and characterizations are required [88]. This observation is not unique to Ambr250 systems though, as data for all processes was spread across the PCA plot for different scales, reflecting the difficulty of matching all conditions across scales. A further 3D plot of the first three PCs (Figure 1S in the Supplementary Information) explaining the highest variance in the original data also resulted in similar conclusions, as most of the variance is explained by the first two principal components.

A 2000 L stainless steel (SS) bioreactor, second only in volume to the 5000 L bioreactors within the dataset, is a second feature of interest in the PCA plot (Figure 3A); this lies outside the confidence interval to the right of the PC1 axis, while other 2000 L bioreactors in this dataset fall within the confidence ellipse. To further investigate why this bioreactor differs, the distribution profiles of the 17 variables listed in Section 2.4 were plotted. The distribution of the complete dataset for four key variables is shown in Figure 3B, namely the impeller diameter, agitation rate, Reynolds number, and P/V, with the complete set of 17 variables provided in Figure 2S in the Supplementary Material.

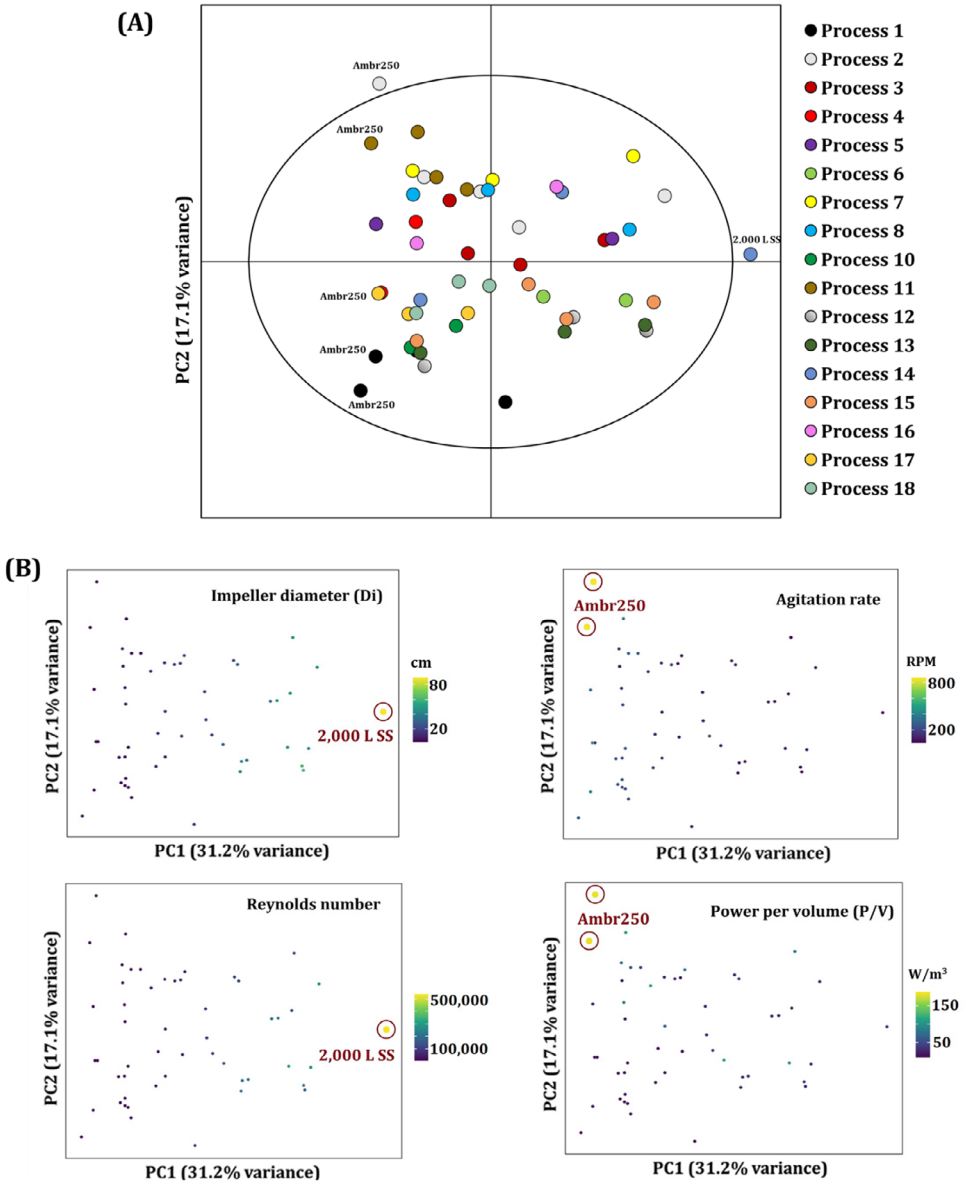

FIGURE 3 Principal component analysis performed to assess similarity between the 18 bioreactor processes. (A) Score plots for the first two principal components (PC1 and PC2). (B) Distribution profiles the plot PC1 versus PC2 for the key variables of impeller diameter, agitation rate, Reynolds number, and *P*/*V*. In (A), multiple dots of the same color indicate the same process. In (B), the color indicator provides a measure of scale and select data points of interest are highlighted with a red circle.

Several variables set the 2000 L SS reactor apart. This bioreactor has by far the highest impeller diameter (indicated by a yellow color) (Figure 3B). The higher Reynolds number also indicates a different flow regime (indicated by yellow) (Figure 3B). Two of the five Ambr250 bioreactors, denoted by the red circles in Figure 3B, have the highest *P*/*V* and agitation rates compared to other bioreactors (indicated by yellow). While it is not always clear why these differences in set up have occurred, given the large public dataset is made up from many different processes, this type of visualization tool is useful to observe variation among data across scales and different bioreactors. One possible reason for the difference in these two Ambr250 bioreactors is that the high power input and agitation in these bioreactors would lead to high dissolved oxygen and mass transfer rates, which would be much greater than when applying the equivalent conditions (i.e., power level) from a larger vessel [89, 90]. This difference also highlights how a constant *P*/*V* criterion applied to scale-up or scale-down is likely to fail, as larger bioreactors typically have lower power requirements.

A further analysis of variation in dynamic cell growth parameters over the duration of the culture, including

time-series data of cell viability, viable cell density, integral viable cell density, as well as lactate concentration was performed by PCA for processes where these data were reported. The results, shown in Figure 3SA in the Supplementary Material, indicate some clear trends in the process data with time, irrespective of the process or bioreactor type, where the general direction of each variable is shown by an arrow. Specifically, changes in PC1, could be linked to IVCD and cell culture duration, as these arrows were more closely aligned with the PC1 axes, whereas changes in PC2 could be linked to cell metabolism and lactate concentration. Changes in viability and VCD with time appeared to affect both PC1 and PC2.

There are some differences between the trajectory of individual processes as a function of time in Figure 3SA, Supplementary Material. These can arise from either different bioreactor geometries or inherent metabolic shifts, including the exhaustion of nutrients and the catabolism of intermediates, including lactate, during the shift from exponential growth to stationary phase. The effect of metabolic shift is particularly clear for the cells in process 10, where the highest net lactate accumulation was observed, compared to other processes and several data points cluster near the edge of the confidence ellipse (Figure 3SA, Supplementary Material).

Several common cellular traits can be assumed across antibody-producing CHO cell processes, however, as a result of most score points in Figure 3SA, Supplementary Material, occurring within the confidence ellipse. PC score data for high CHO cell viability (in yellow), for example, is clustered in the left upper quadrant (Figure 3SB, Supplementary Material), which is diagonally opposed to high culture duration that occurs in the bottom right quadrant (in yellow) (Figure 3SB, Supplementary Material). These data indicate that across the processes, cell viability was generally high (often greater than 95%) early in the process, with some decline in viability in the later stages of cultivation. Most CHO-based processes, however, displayed an extended viability over the course of cultivation, as shown by the spread of yellow datapoints. The VCD follows an interesting pattern, with the PC score data increasing from the middle left (purple) to the top right (yellow), indicating the transition from cell growth to other metabolic states, which can be attributed to specific culture properties, such as nutrient depletion and waste accumulation. The concentration of lactate is an important process indicator that is typically measured and monitored in CHO cell culture [91, 92]. A nonmonotonic trend is presented in the PCA plot shown in Figure 3SB, Supplementary Material, where the pattern moves from purple on the left to green and yellow in the bottom middle and then back to purple on the right. This pattern, observed for all CHO-based processes, may arise due to variations in lactate production and consumption. The extended culture viability and enhanced VCD result in an increase in IVCD over the process, which leads to a similar distribution profile for both the IVCD and the culture duration (Figure 3SB, Supplementary Material).

The PCA plots indicate that horizontal knowledge transfer may be possible across processes, that is, knowledge about one process may be transferred to another process, as we expect that data acquired for a new antibody product obtained from a different CHO cell expression system are most likely to be located within the same confidence ellipse with similar trajectories in the PCA space. This hypothesis was further tested by in-house data at various scales (5, 200, and 2000 L). The results (not shown) verified that all the observations for the new data fell within the confidence ellipse over the culture evolution.

3.4 | Prediction performance of ML models

A range of machine learning models based on ANN and XGBoost algorithms were trained and tested for three process examples (i.e., process 1, 2, and 3) selected from the dataset of 18 different processes and applied to three different case studies, as described in Table 1 and Section 2.5, with the purpose of testing whether the model: (1) could predict a new unseen process, (2) could predict across scales when provided with only small-scale data; and (3) whether the addition of entity embedding improved model prediction. The effect of one-hot encoding on model predictions was also tested. The optimized hyperparameters for these models are presented in Table 2S in the Supplementary Material.

The performance of the ANN models for processes 1, 2, and 3 was evaluated on both test data and unseen data, as detailed in Section 2.5.3, generating the predicted VCD shown in Figure 4. A good predictive model will generate data that clusters close to the y = x line—that is, there will be a good match between the predicted VCD (the vertical axis) and the corresponding actual VCD (the horizontal axis).

The ANN models for the first case study worked well on the test data for all three processes (right graph for each of the three processes) but less well on the unseen data (left graph for each process), which is expected. The basic ANN model generated VCD data that was scattered about the x = y line (black series) and resulted in a higher root mean square error (RMSE) for the unseen data compared to the test data (purple and yellow bars).

The inclusion of small-scale data in case study 2, where the VCD for the same process was predicted at larger scale, generally improved model predictions, as shown by the

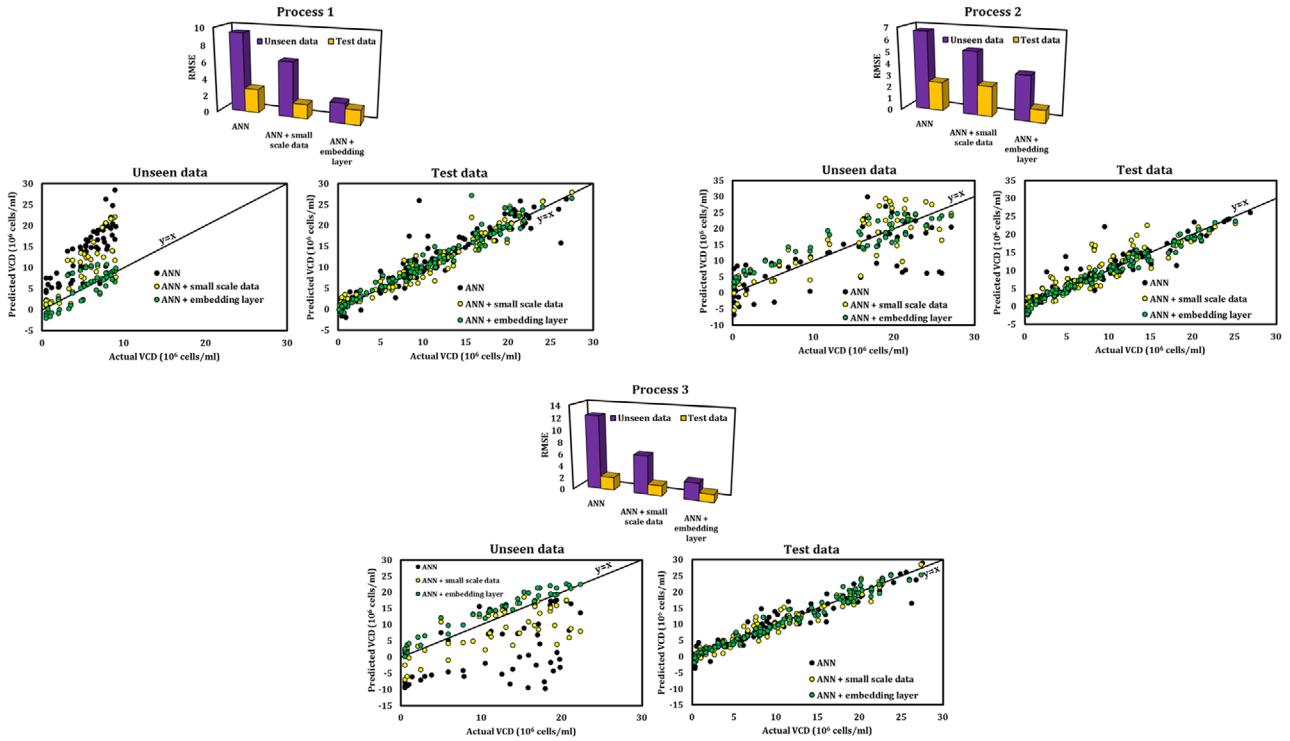

FIGURE 4 The prediction of VCD by the ANN models developed in this study for three example processes (process 1, 2, and 3) based on the three case studies, as shown in Table 1. The bar charts indicate the root mean square error (RMSE) for all three cases studies (left to right) for each of the three processes.

yellow data series and improved RMSE for the unseen data for each process (purple bar for each process).

The greatest improvement was observed in case study three, with the addition of the embedding layer to the ANN model, which resulted in the data clustering more closely to the x = y line (green series) and a lower RMSE for the unseen data for all three processes (purple bar for each process). This outcome indicates that creation of a D-dimensional vector that contains numerical values encoding each process is necessary to capture possible relationships among processes in the artificial neural network model.

A comparison between the predictive performance of different XGBoost models developed is shown in Figure 5. Similar to the ANN predictions, the XGBoost models had a lower ability to predict a new scaling process in case study one, where the basic XGBoost model was tested, resulting in a greater RMSE value for the unseen data (purple bar for each process) than the test data (yellow bar for each process) for all three processes (Figure 5). The magnitude of the RMSE error obtained by XGBoost for the unseen datasets, however, was lower than that obtained using the ANN models, with an RMSE of 5.8, 2.8, and 3.8 for processes 1–3 compared to 9.4, 6.7, and 12.3 for the ANN models respectively. The inclusion of small-scale data for case study 2, improved two of the three predictions of VCD, with a small decrease in the RMSE for the unseen larger scale data (purple bar) for these processes (process 1 and 2; Figure 5).

Unlike the ANN with the embedding layer, the one-hot encoding representation of the categorial variable (i.e., the process type) did not improve the predictive performance of the XGBoost model in case study three. This could be due to the inefficiency of one-hot encoding for the high cardinality categorical variable, where the cardinality refers to the number of values that can be assigned to the categorical variable, that is, the 18 different processes. One-hot encoding generates 18 new features, leading to a significant increase in the dimensionality of the feature representations. Furthermore, the one-hot encoding is unlike entity embedding, which maps the categorical variable into a D-dimension numerical space determined in hyperparameter tuning that conveys the intrinsic properties of each process and can consider possible similarities between processes. Instead, the one-hot encoding representation of the categorical variable cannot include the relationship between processes [93]. Despite these drawbacks, the processes were better predicted by the XGBoost models, with lower RMSE. The predictive performance of either the ANN models or XGBoost models was also shown to be highly process-specific, given the varying RMSE values calculated for each of the three processes.

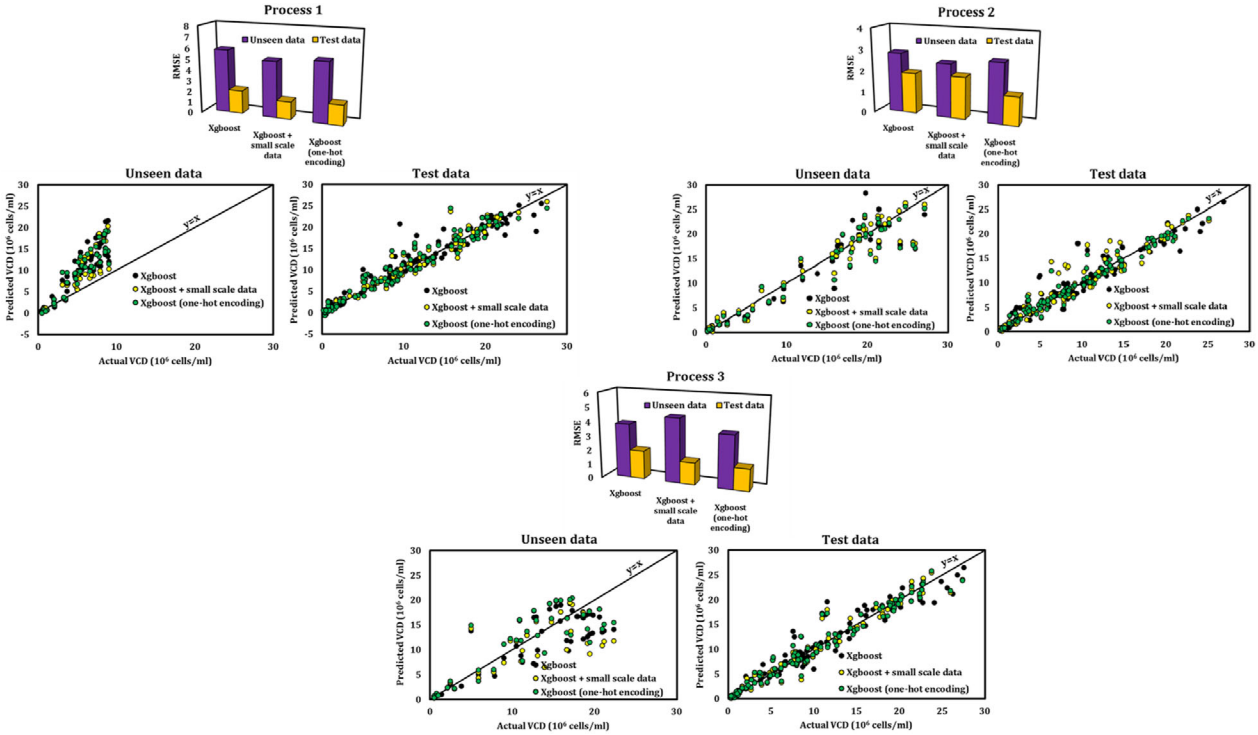

FIGURE 5 The prediction of VCD by the XGBoost models developed in this study for three example processes (process 1, 2, and 3) based on the three case studies shown in Table 1. The bar charts indicate the root mean square error (RMSE) for all three cases studies (left to right) for each of the three processes.

3.5 | Potential application and current limitations of ML models for predicting process performance

The proof-of-concept ML models demonstrated with test cases 2 and 3 above, either with ANN or XGBoost components, could potentially be applied to predict process performance across scales during process development. Application would require a new training dataset of mammalian production processes similar to the process of interest undergoing scaling, such as historical scaling data collected for industrial processes. Data would also need to be collected for the new process of interest at small scale, which typically occurs early in process development.

One of the limitations of the dataset collated and examined here is that engineering features of the bioreactor geometry, such as the vessel diameter and the aspect ratio, only vary between scales and remain constant for processes reported at the same scale. This results in low variability in scale-sensitive factors, which limits the capability of the models to predict the effect of scaling parameters on VCD output. Consequently, culture duration was identified as the top predictor of output in an analysis of feature importance by the SHAP method (shown in Figure 4S in the Supplementary Information).

An ideal training dataset would include a much broader set of process data at each scale, including different bioreactor characteristics including agitation, power and aeration conditions, mixing times, and culture volumes (e.g., some of the features considered within Figure 4S, Supplementary Material); this dataset would also ideally include data for different bioreactor geometries, such as different vessel diameters and aspect ratios, at each scale and could also include other characteristics of the chosen cell line. The development of an ideal training set with adequate process variation and similarity to the process of interest undergoing scaling is expected to further reduce error beyond the reductions observed here.

A further extension would also be to consider other ML algorithms. While ANN and boosting algorithms, such as XGBoost, are among the most powerful models for predicting nonlinear, complex systems, various other ML algorithms, such as support vector machines, Gaussian processes and random forests, to name a few, could also be tested for similar application to scaling.

3.6 | Extension to predict scaling parameters

In an extension of the models presented in Section 3.4, we next developed a new ML model to predict popular scaling criteria used by industry to scale bioreactors. P/V

and the oxygen mass transfer coefficient (k_La) or oxygen transfer rate (OTR) are three commonly used scaling criteria. Matching P/V is believed to ensure comparable shear stress, mixing conditions, and oxygen transfer rates between two scales.

In this example, we set out to determine a P/V that would be needed to achieve a target peak VCD, starting from a given seeding density within a bioreactor. This scenario is often encountered when a target peak VCD is known at smaller scale and the process of scaling seeks to achieve the same target peak VCD at larger scale. P/V was chosen for demonstration, as these data were available within the public dataset compiled here, while oxygen transfer data were constrained (see Section 2.3). The first step involved development of a function for P/V and bioreactor or process variables, where dimensionality reduction was used to reduce the number of variables and multicollinearity. Bioreactor and process variables considered included volume, impeller diameter, impeller tip speed, Re, mixing time and agitation rate, which were found to be highly correlated with the vessel diameter, with correlation coefficients between 0.6 to more than 0.9, indicated by the red color in Figure 5S in the Supplementary Information. Significant multicollinearity was also identified among variables, as indicated by the large VIF values listed in Table 3S, Supplementary Material. Dimensionality reduction was used to largely eliminate this multicollinearity, resulting in VIF values close to one, as shown in Table 3S, Supplementary Material. The choice of variables was based on the results obtained from the distance correlation (Figure 5S, Supplementary Material) (e.g., out of several variables that were found to have a strong correlation with vessel diameter, as exemplified above, vessel diameter was retained as a feature, since it is readily available). This was followed by checking the VIF values for the reduced feature space. The final function with reduced dimensions can be summarized as follows:

$$\mathrm{P/V} = \mathrm{f}(\text{seeding density},\ \text{peak VCD},\ \text{aspect ratio},\ \text{vessel diameter},\ \mathrm{Di/Dt})$$

A CatBoost [94] machine learning model was developed to predict P/V, as this approach can be applied to small datasets; in addition, this method is able to directly incorporate high cardinality categorical variables (i.e., allows unique values) without increasing the dimension of the dataset; this is a great advantage for the current dataset with 18 different process names (each of which are categorical variables with significant cardinality, i.e., different process names). Moreover, a CatBoost algorithm can provide an estimate of uncertainty associated with each prediction.

For each of the 18 processes in the dataset, a specific model was trained using small-scale data and used to predict the P/V value at a larger scale for that specific process. Large-scale observations were excluded from the training to allow the model to be tested, while the remaining small-scale data from that process plus the data obtained from the 17 other processes were used to train the model. For each process model, the hyperparameters including depth, maximum number of trees (num_boost_round), and learning rate were then tuned by Bayesian optimization in Python, resulting in a total of 18 optimized models (Table 4S, Supplementary Material).

This approach gave good predictions of P/V, close to the actual P/V values for the larger bioreactor (the target scale within each process), while potentially achieving the same peak VCD as occurred in the smaller bioreactor (the reference scale), as shown in Figure 6. The uncertainty estimated by the CatBoost algorithm via a virtual ensemble of models (where the ensemble count was 50) [94] is also represented by the error bars in this figure.

Good agreement was observed for most processes. Nevertheless, for some processes, in particular process 4, the discrepancy between the predicted and actual values is significant. This could be due to a much greater power input used in process 4 at larger scale, which may have been used to maintain the dissolved oxygen at the process setpoint over the culture duration and to achieve a sufficient homogenization, as stated by the manufacturing team [95].

This proof-of-concept approach demonstrates the possibility of developing ML models for the estimation of scaling parameters and future work could seek to use ML predicted parameters to match bioreactor performance across scales. Application of this approach would again require a good historical or experimental dataset of bioprocesses similar to the target process across two or more scales. This could include data for the features selected here, that is, seeding density, peak VCD, aspect ratio, vessel diameter, and Di/Dt or other combinations of variables similarly selected to reduce dimensions and VIF, as described for the function developed above. The data could then be used to create data-driven models that predict key scale-dependent factors, such as P/V, or with extension OTR, at both reference and target scales. Further, optimization algorithms might be developed to find suitable agitation and aeration values, to achieve equivalent performance between two scales, given cell line characteristics and limitations, such as shear stress. In this way, equivalent bioreactor performance, such as equal peak VCD or comparable product concentrations, could potentially be achieved between reference and target scales.

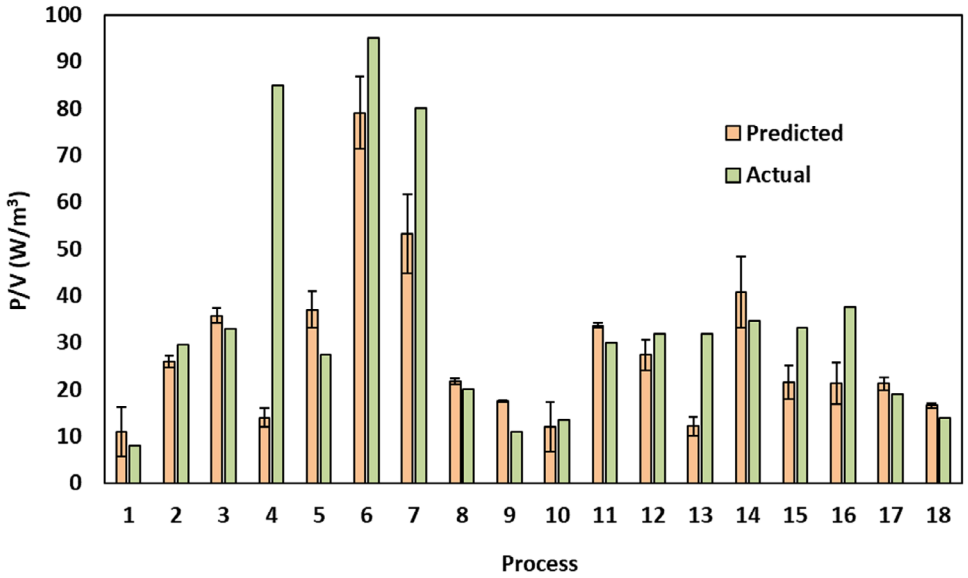

FIGURE 6 The predicted power inputs (P/V) from the CatBoost model at the larger target scale and actual P/V measurements reported experimentally for each process. The error bars represent the uncertainty from a virtual ensemble of 50 models.

4 | CONCLUSIONS

One of the most difficult aspects of transferring cell cultures from one scale to another is the determination of suitable scaling factors that can give reproducible culture performance between scales. Factors affecting scaling were reviewed and the present modeling framework demonstrated the potential of machine learning tools to build mathematical relationships among important engineering scale-dependent factors that can predict cell growth. The unsupervised learning methods, based on principal component analysis, identified differences and similarities between bioreactors across 18 different mAb production processes involving CHO cells. The wide range of scale-dependent factors, arising from different design configurations, was one of the main sources of difference between small- and large-scale bioreactors; for example, very high Reynolds numbers were observed at large scales and very high agitation rates at small scales, such as Ambr250 bioreactors. The similarity between growth-related factors in these CHO cell culture-based processes, however, underscored the potential for knowledge transfer between processes. The supervised learning models based on ANN could predict process performance for some processes and had higher predictive capability when process characteristics were incorporated using entity embeddings. The CatBoost algorithm was also efficient in handling the high cardinality categorical features in the small dataset, using its in-built target encoding capability. The need for comprehensive scaling datasets was identified, to build more powerful, generalized models capable of predicting suitable scaling factors. These should ideally include sufficient variability generated under a wide range of operating conditions, including gassing flow rates and agitation rates, with data collected for the same process at different scales and data collected across processes at the same scale, which would allow both vertical and horizontal knowledge transfer, respectively. These approaches may be useful for future investigations using ML for scale-up of bioreactors with equivalent performance.

AUTHOR CONTRIBUTIONS

M. Karimi Alavijeh: Conceptualization; Methodology; Formal analysis; Investigation; Visualization; Software; Writing—Original Draft. **Y.Y. Lee**: Project administration; Supervision; Writing—Review & Editing. **S.L. Gras**: Conceptualization; Supervision; Resources; Project administration; Writing—Review & Editing.

ACKNOWLEDGMENTS

The Victorian Government provided financial support for this research under the Higher Education State Investment Fund (VHESIF) program, with additional support from CSL Innovation, Australia.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Sally L. Gras https://orcid.org/0000-0002-4660-1245

REFERENCES

1. Ecker DM, Crawford TJ, Seymour P. The therapeutic monoclonal antibody product market. Bioprocess Int. 2020. https://www.bioprocessintl.com/economics/the-therapeutic-monoclonal-antibody-product-market
2. Karimi Alavijeh M, Baker I, Lee YY, Gras SL. Digitally enabled approaches for the scale up of mammalian cell bioreactors. *Digital Chem. Eng*. 2022;4:100040.
3. Xu S, Hoshan L, Jiang R, et al. A practical approach in bioreactor scale-up and process transfer using a combination of constant P/V and vvm as the criterion. *Biotechnol Progr*. 2017;33:1146-1159.
4. Li F, Hashimura Y, Pendleton R, Harms J, Collins E, Lee B. A systematic approach for scale-down model development and characterization of commercial cell culture processes. *Biotechnol Progr*. 2006;22:696-703.
5. Chaudhary G, Luo R, George M, Tescione L, Khetan A, Lin H. Understanding the effect of high gas entrance velocity on Chinese hamster ovary (CHO) cell culture performance and its implications on bioreactor scale-up and sparger design. *Biotechnol Bioeng*. 2020;117:1684-1695.
6. Anane E, Knudsen IM, Wilson GC. Scale-down cultivation in mammalian cell bioreactors—the effect of bioreactor mixing time on the response of CHO cells to dissolved oxygen gradients. *Biochem Eng J*. 2021;166:107870.
7. Ben Yahia B, Malphettes L, Heinzle E. Macroscopic modeling of mammalian cell growth and metabolism. *Appl Microbiol Biotechnol*. 2015;99:7009-7024.
8. Goudar CT. Computer programs for modeling mammalian cell batch and fed-batch cultures using logistic equations. *Cytotechnology*. 2012;64:465-475.
9. Robitaille J, Chen J, Jolicoeur M. A single dynamic metabolic model can describe mAb producing CHO cell batch and fed-batch cultures on different culture media. *PLoS One*. 2015;10:e0136815.
10. Xing Z, Bishop N, Leister K, Li ZJ. Modeling kinetics of a large-scale fed-batch CHO cell culture by Markov chain Monte Carlo method. *Biotechnol Progr*. 2010;26:208-219.
11. O'Brien CM, Zhang Q, Daoutidis P, Hu W-S. A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. *Metab Eng*. 2021;66:31-40.
12. Villiger TK, Neunstoecklin B, Karst DJ, et al. Experimental and CFD physical characterization of animal cell bioreactors: from micro- to production scale. *Biochem Eng J*. 2018;131:84-94.
13. Thomas JA, Liu X, DeVincentis B, et al. A mechanistic approach for predicting mass transfer in bioreactors. *Chem Eng Sci*. 2021;237:116538.
14. Scully J, Considine LB, Smith MT, et al. Beyond heuristics: cFD-based novel multiparameter scale-up for geometrically disparate bioreactors demonstrated at industrial 2kL–10kL scales. *Biotechnol Bioeng*. 2020;117:1710-1723.
15. Delafosse A, Collignon M-L, Calvo S, et al. CFD-based compartment model for description of mixing in bioreactors. *Chem Eng Sci*. 2014;106:76-85.
16. Pigou M, Morchain J. Investigating the interactions between physical and biological heterogeneities in bioreactors using compartment, population balance and metabolic models. *Chem Eng Sci*. 2015;126:267-282.
17. Cytiva Life Sciences. Engineering Characterization of the Single-Use Xcellerex XDR-50 Stirred-Tank Bioreactor System; 2020.
18. Cytiva Life Sciences. Engineering Characterization of the Single-Use Xcellerex XDR-200 Stirred-Tank Bioreactor System; 2020.
19. Cytiva Life Sciences. Engineering Characterization of the Single-Use Xcellerex XDR-1000 Stirred-Tank Bioreactor System; 2020.
20. Cytiva Life Sciences. Xcellerex XDR Cell Culture Bioreactor Systems; 2020.
21. Eppendorf. Single-Use Simplicity, BioBLU® c and BioBLU p Single-Use Vessels for Cell Culture; 2022.
22. Han Xiaofeng, Willard S, Sha M, Cell Culture Scale-Up in BioBLU® c Rigid-Wall, Single-Use Bioreactors; 2016.
23. Merck Millipore. Scalability of the Mobius® Single-use Bioreactors; 2016.
24. Thermo Fisher Scientific. 50 L HyPerforma DynaDrive Single-Use Bioreactor; 2021.
25. Thermo Fisher Scientific. 500 L HyPerforma DynaDrive Single-Use Bioreactor; 2021.
26. Thermo Fisher Scientific. 3,000L and 5,000L HyPerforma DynaDrive Single-Use Bioreactor; 2021.
27. Sartorius. UniVessel® Glass Reliability and Continuity; 2018.
28. Dreher T, Husemann U, Adams T, de Wilde D, Greller G. Design space definition for a stirred single-use bioreactor family from 50 to 2000 L scale. *Eng Life Sci*. 2014;14:304-310.
29. BioProcess International. Innovations in Cell Culture; 2014.
30. Pall Corporation. Cultivation of Chinese Hamster Ovary (CHO) Cells in Allegro™ STR 1000 Single-Use Stirred Tank Bioreactor System; 2016.
31. Pall Corporation. Scalability Between the Allegro™ STR 50 and STR 500 Bioreactors in a CHO-S Fed-Batch Process; 2020.
32. Applikon Biotechnology. The Applikon 2 - 7 liter Autoclavable Bioreactors; 1994.
33. Applikon Biotechnology. Glass Autoclavable Bioreactors, the World Wide Standard; 2021.
34. Antonakoudis A, Strain B, Barbosa R, Jimenez del Val I, 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.
35. Roy S, Sharma P, Nath K, Bhattacharyya DK, Kalita JK. Preprocessing: a data preparation step. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, eds. *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press; 2019:463-471.
36. Chio C, Freeman D. *Machine Learning and Security: Protecting Systems with Data and Algorithms*. O'Reilly Media; 2018.
37. Hoehse M, Alves-Rausch J, Prediger A, Roch P. Near-infrared spectroscopy in upstream bioprocesse. *Pharmaceutical Bioprocessing*. 2015;3:153-172.
38. Smiatek J, Clemens C, Herrera LM, et al. Generic and specific recurrent neural network models: applications for large and small scale biopharmaceutical upstream processes. *Biotechnology Rep*. 2021;31:e00640.
39. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw*. 2008;25:1-18.
40. Weiner J. Package 'pca3d': Three Dimensional PCA Plots; 2020. https://cran.r-project.org/web/packages/pca3d/pca3d.pdf
41. The CMC Biotech Working Group. A-Mab: a Case Study In Process Development; 2009. https://ispe.org/publications/guidance-documents/a-mab-case-study-in-bioprocess-developments

42. He C, Ye P, Wang H, Liu X, Li F. A systematic mass-transfer modeling approach for mammalian cell culture bioreactor scale-up. *Biochem Eng J*. 2019;141:173-181.
43. Kotidis P, Kontoravdi C. Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab Eng Commun*. 2020;10:e00131.
44. Hall LM, Hill DW, Menikarachchi LC, Chen M-H, Hall LH, Grant DF. Optimizing artificial neural network models for metabolomics and systems biology: an example using HPLC retention index data. *Bioanalysis*. 2015;7:939-955.
45. Wang S, Fan K, Luo N, et al. Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nat Commun*. 2019;10:4354.
46. Xu C, Jackson SA. Machine learning and complex biological data. *Genome Biol*. 2019;20:76.
47. Nogueira F. Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python; 2014. https://github.com/fmfn/BayesianOptimization
48. Passos D, Mishra P. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemom Intell Lab Syst*. 2022;223:104520.
49. Hutter C, von Stosch M, Cruz Bournazou MN, Butté A. Knowledge transfer across cell lines using hybrid Gaussian process models with entity embedding vectors. *Biotechnol Bioeng*. 2021;118:4389-4401.
50. Brandt S, Sittel F, Ernst M, Stock G. Machine learning of biomolecular reaction coordinates. *J Phys Chem Lett*. 2018;9:2144-2150.
51. Van Camp P-J, Haslam DB, Porollo A. Prediction of antimicrobial resistance in gram-negative bacteria from whole-genome sequencing data. *Front Microbiol*. 2020;11:1013.
52. Tokuyama K, Shimodaira Y, Kodama Y, et al. Soft-sensor development for monitoring the lysine fermentation process. *J Biosci Bioeng*. 2021;132:183-189.
53. Shi S, Xu G. Identification of phosphorus fractions of biofilm sludge and phosphorus release, transformation and modeling in biofilm sludge treatment related to pH. *Chem Eng J*. 2019;369:694-704.
54. Deepthi K, Jereesh AS, Yuansheng Liu. A deep learning ensemble approach to prioritize antiviral drugs against novel coronavirus SARS-CoV-2 for COVID-19 drug repurposing. *Appl Soft Comput*. 2021;113:107945.
55. Magar R, Yadav P, Barati Farimani A. Potential neutralizing antibodies discovered for novel corona virus using machine learning. *Sci Rep*. 2021;11:5261.
56. Chen T, Guestrin C. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:785-794.
57. Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Statist*. 2007;35:2769-2794.
58. Forthofer RN, Lee ES, Hernandez M. 13 - linear regression. In: Forthofer RN, Lee ES, Hernandez M, eds. Biostatistics. *2nd ed*. Academic Press; 2007:349-386.
59. Lundberg SM, Lee S-I. *Advances in Neural Information Processing Systems*. Curran Associates, Inc; 2017.
60. Baptista ML, Goebel K, Henriques EMP. Relation between prognostics predictor evaluation metrics and local interpretability SHAP values. *Artif Intell*. 2022;306:103667.
61. Meena J, Hasija Y. Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. *Comput Biol Med*. 2022;146:105505.
62. Ahmed SK, Antoniou C, Guenard R, Romero-Torres S. Hybrid model identification for monoclonal antibody production bioreactor– a digital twin. *Am Pharma Rev*. 2019;22. https://www.americanpharmaceuticalreview.com/Featured-Articles/517739-Hybrid-Model-Identification-for-Monoclonal-Antibody-Production-Bioreactor-A-Digital-Twin/
63. Junker BH. Scale-up methodologies for Escherichia coli and yeast fermentation processes. *J Biosci Bioeng*. 2004;97:347-364.
64. Xu S, Bowers J, Seamans TC, Nyberg G. Bioreactor scale-up. In: Kirk-Othmer, ed. *Encyclopedia of Chemical Technology*, 2018:1-35.
65. Ju LK, Chase GG. Improved scale-up strategies of bioreactors. *Bioprocess Eng*. 1992;8:49-53.
66. Wang P, Wang S, Gu Y, Si Q, Yuan S. The effect of the cavity formation on the energy consumption characteristics of the agitated gas–liquid bioreactor. *AIP Adv*. 2022;12:015103.
67. Grutzmacher D, Proquip. 2019.
68. Bates RL, Fondy PL, Corpstein RR. Examination of some geometric parameters of impeller power. *Ind Eng Chem Process Des Dev*. 1963;2:310-314.
69. Koerich DM, Rosa LM. Optimization of bioreactor operating conditions using computational fluid dynamics techniques. *Can J Chem Eng*. 2017;95:199-204.
70. Nienow AW. Stirring and stirred-tank reactors. *Chem Ing Tech*. 2014;86:2063-2074.
71. van 't Riet K, van der Lans RGJM. 2.07 - mixing in bioreactor vessels. In: Moo-Young M, ed. Comprehensive Biotechnology. *2nd ed*. Academic Press; 2011:63-80.
72. Varley J, Birch J. Reactor design for large scale suspension animal cell culture. *Cytotechnology*. 1999;29:177.
73. Khan M, Watford H. Bioreactor for the cultivation of mammalian cells, Lonza Biologics plc, Patent number:US10883076B2; 2018.
74. Zhong JJ. 2.14 - bioreactor engineering. In: Moo-Young M, ed. Comprehensive Biotechnology. *2nd ed*. Academic Press; 2011:165-177.
75. Marks DM. Equipment design considerations for large scale cell culture. *Cytotechnology*. 2003;42:21-33.
76. Doran PM. Chapter 8 - mixing. In: Doran PM, ed. Bioprocess Engineering Principles. *2nd ed*. Academic Press; 2013:255-332.
77. The CMC Biotech Working Group. A-Mab: a Case Study in Process Development; 2009.
78. Kaiser SC, Löffelholz C, Werner S, Eibl D. CFD for characterizing standard and single-use stirred cell culture bioreactors. In: Minin IV, Minin OV, eds. *Computational Fluid Dynamics Technologies and Applications*. IntechOpen; 2011.
79. Sieblist C, Jenzsch M, Pohlscheidt M. Equipment characterization to mitigate risks during transfers of cell culture manufacturing processes. *Cytotechnology*. 2016;68:1381-1401.
80. Kubera P. Testing and simulation approaches for single-use bioreactor scale-up. *Pharm Technol*. 2017;41:42-45.
81. Srivastava VC, Mishra IM, Suresh S. 2.69 - oxygen mass transfer in bioreactors. In: Moo-Young M, ed. Comprehensive Biotechnology. *2nd ed.* Academic Press; 2011:947-956.
82. Nienow AW. Reactor engineering in large scale animal cell culture. *Cytotechnology*. 2006;50:9.

83. Ebrahimi M, Tamer M, Villegas RM, Chiappetta A, Ein-Mozaffari F. Application of CFD to analyze the hydrodynamic behaviour of a bioreactor with a double impeller. *Processes*. 2019;7:694.
84. Manahan M, Nelson M, Cacciatore JJ, Weng J, Xu S, Pollard J. Scale-down model qualification of ambr® 250 high-throughput mini-bioreactor system for two commercial-scale mAb processes. *Biotechnol Progr*. 2019;35:e2870.
85. Zhang X, Moroney J, Hoshan L, Jiang R, Xu S. Systematic evaluation of high-throughput scale-down models for single-use bioreactors (SUB) using volumetric gas flow rate as the criterion. *Biochem Eng J*. 2019;151:107307.
86. Bareither R, Bargh N, Oakeshott R, Watts K, Pollard D. Automated disposable small scale reactor for high throughput bioprocess development: a proof of concept study. *Biotechnol Bioeng*. 2013;110:3126-3138.
87. Kwan B, Bowers J, Chauhan G, Bandyopadhyay A, Ling WL. *2020 Virtual AIChE Annual Meeting*. AIChE; 2020.
88. Sandner V, Pybus LP, McCreath G, Glassey J. Scale-Down Model development in ambr systems: an Industrial Perspective. *Biotechnol J*. 2019;14:1700766.
89. Kistler C, Pollard J, Ng LS, Streefland M. High-throughput bioprocess development. *Genetic Engineering & Biotechnology News*. 2016;36:30-31.
90. Xu P, Clark C, Ryder T. Characterization of TAP Ambr 250 disposable bioreactors, as a reliable scale-down model for biologics process development. *Biotechnol Progr*. 2017;33:478-489.
91. Zagari F, Jordan M, Stettler M, Broly H, Wurm FM. Lactate metabolism shift in CHO cell culture: the role of mitochondrial oxidative activity. *New Biotechnol*. 2013;30:238-245.
92. Buchsteiner M, Quek L-E, Gray P, Nielsen LK. Improving culture performance and antibody production in CHO cell culture processes by reducing the Warburg effect. *Biotechnol Bioeng*. 2018;115:2315-2327.
93. Guo C, Berkhahn F. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*. 2016. https://doi.org/10.48550/arXiv.1604.06737
94. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inform Process Syst*. 2018.
95. Janke F, Kober L, Glaser R. Scale-Up of a Biosimilar Production Process with CHO Cells from Small to Bench Scale; 2021.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Karimi Alavijeh M, Lee YY, Gras SL. A perspective-driven and technical evaluation of machine learning in bioreactor scale-up: A case-study for potential model developments. *Eng Life Sci*. 2024;24:e2400023. https://doi.org/10.1002/elsc.202400023