Biochemical Engineering Journal 185 (2022) 108499

Contents lists available at ScienceDirect

Biochemical Engineering Journal

journal homepage: www.elsevier.com/locate/bej

# Data-driven modelling for resource recovery: Data volume, variability, and visualisation for an industrial bioprocess

Oliver J. Fisher[a], Nicholas J. Watson[a], Laura Porcu[b,c], Darren Bacon[b], Martin Rigley[b], Rachel L. Gomes[a,*]

[a] Food Water Waste Research Group, Faculty of Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom
[b] Lindhurst Engineering Ltd., Midland Road, Sutton in Ashfield, Nottinghamshire NG17 5GS, United Kingdom
[c] Energy Innovation & Collaboration, University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, United Kingdom

ARTICLE INFO

*Keywords:*
Data-driven models
Bioprocess
Data volume
Data variability
Data visualisation

ABSTRACT

Advances in industrial digital technologies have led to an increasing volume of data generated from industrial bioprocesses, which can be utilised within data-driven models (DDM). However, data volume and variability complications make developing models that captures the underlying biological nature of the bioprocesses challenging. In this study, a framework for developing data-driven models of bioprocesses is proposed and evaluated by modelling an industrial bioprocess, which treats industrial or agrifood wastewaters whilst simultaneously generating bioenergy. Six models were developed to predict the reduction in chemical oxygen demand from the wastewater by the bioprocess and statistically evaluated using both testing data (randomly partitioned data from the model development) and unseen data (new data not used during the model development). The statistical error metrics employed were the coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The stacked neural network model was best able to model the bioprocess, having the highest accuracy on the testing data ($R^2$: 0.98; RMSE: 1.29; MAE: 2.27; MAPE: 4.08) and the unseen data ($R^2$: 0.82; RMSE: 2.57; MAE: 1.75; MAPE: 3.68). Data visualisation is used to observe (or confirm) whether new data points are within the model boundaries, helping to increase confidence in the model's predictions on future data.

## 1. Introduction

Bioprocesses have complex dynamics and are subject to disturbances, which makes modelling bioprocesses challenging yet necessary for process understanding, model-based optimisation and scale-up [1]. The biological nature of bioprocesses makes constructing models based on physical laws particularly challenging [2]. An alternative are data-driven models (DDMs) that fit process data to algorithms to discover knowledge about a system and/or make predictions about the system [3]. However, for a DDM to be reflective of a bioprocess it requires the modelling data to be representative of the underlying biological nature of the processes [4]. This raises questions from a data volume and data variability perspective, as well as how to increase trust in the DDM. To demonstrate how these questions may be addressed and how the data and DDM results can be visualised to improve communication and trustworthiness, this work develops a DDM of an industrial bioprocess. The bioprocess is capable of treating a variety of wastewaters, from sources including agriculture, brewing, soft drinks, foods, bio-manufacture residues, to reduce the pollutant load and improve the water quality for reuse, whilst simultaneously generating bioenergy. Predictions from the DDM may be utilised to analyse the effects that varying wastewater characteristics and process conditions have on the bioprocess' ability to remediate chemical oxygen demand (COD), a key water quality parameter.

*Abbreviation:* ANN, Artificial Neural Network; BES, Bioelectrochemical Systems; COD, Chemical Oxygen Demand; R2, Coefficient of Determination; Cond., Conductivity; DDM, Data-Drive Model; GP, Gaussian Process; HRT, Hydraulic Residence Time; MEA, Mean Average Error; MAPE, Mean Average Percentage Error; NCA, Neighbourhood Component Analysis; PCP, Parallel Coordinate Plot; PCA, Principal Component Analysis; RF, Random Forest; RMSE, Root Mean Squared Error; SVR, Support Vector Regression; P, System Pressure; T, System Temperature; TC, Total Carbon; TIC, Total Inorganic Carbon; tN, Total Nitrogen; TOC, Total Organic Carbon; TSS, Total Suspended Solids; WH, Water Hardness; WQA, Water Quality Analysis.

* Corresponding author.
*E-mail address:* rachel.gomes@nottingham.ac.uk (R.L. Gomes).

https://doi.org/10.1016/j.bej.2022.108499
Received 12 January 2022; Received in revised form 9 May 2022; Accepted 27 May 2022
Available online 30 May 2022
1369-703X/

There is no single definition of a bioprocess. For this work, a bioprocess is defined as input streams passing through a bioprocess that changes their nature into output streams, which may comprise multiple products and/or waste materials. In order for a DDM to be representative of a bioprocess, the model data statistics (mean, $\bar{x}$; standard deviation, s; sample range, etc.) should be as similar as possible to the bioprocess parameters (mean, μ; standard deviation, $\sigma$, bioprocess range, etc.) [5]. Often the bioprocess parameters are not known but the volume of data is considered sufficient to assume that the model data statistics equate to the bioprocess parameters [6]. This assumption fails if the data does not represent the whole system due to temporal influences being limited, so variability in the data is not fully realised. Furthermore, there are sectors where data generation and collection are inherently limited. For example, in collecting the data to model the industrial bioprocess, samples of the bioprocesses input and output wastewater streams were collected once a week and water quality analysis was conducted to characterise them. A year's worth of sampling would only produce 52 data points, which is several magnitudes smaller than the 10,000 s of datapoints produced in the manufacturer of semiconductors [7].

Process engineering systems contain inherent variation (e.g. fluctuations in process temperatures, pressure, and flowrates, human operators) [8]. Bioprocesses contain additional variability as their feedstocks can be subjected to external variability (e.g. changes in supplier, local agronomic conditions, cultivation and harvesting practises, seasonal, and storage and transportation) [8]. Furthermore, a high level of variability is often present in processes that use wastes are feedstock and/or processes, especially if there is a biological component. This will become more relevant as society transitions towards a circular economy using waste as feedstock, because wastes are inherently more variable than traditional feedstocks [9,10]. This is true of the bioprocess case study, whose efficiency at reducing pollutant load in the wastewater and generating bioenergy varies with feedstock characteristics fluctuations (temperature, pH, composition, etc.). As a bioprocess's variability increases the likelihood that the model data accurately represents the bioprocess decreases. Exploratory data analysis and statistical inference are techniques to investigate the model data and determine if they are representative of the bioprocess [5]. When the bioprocess parameters are unknown, these techniques can be used to provide estimates and to establish the boundaries of the model. In addition to understanding how data volume and variability informs effective modelling, this work also details how data visualisation techniques may be used to visualise the model boundaries and improve trust in which regions of the bioprocess the DDM accurately represents.

Existing frameworks for modelling bioprocesses frequently include an experimental design stage to ensure the collection of data that contains information on the process dynamics [2,11–14]. The experimental design stage can be improved using design of experiments methods to determine the relationship between bioprocess parameters and outputs [11,14]. Although, these methods rely on prior knowledge of the parameters' boundary values, which are often unknown and require additional experimental work to discover. Recent work has proposed methods to reduce the number of experiments required [12,15]. However, while experimental design has proved an effective method to collect data that captures the process dynamics of a bioprocess, the cost from the disruption to normal operating conditions may make any experimental work unfeasible when modelling industrial bioprocesses. Instead, alternative methods are necessary to ensure the data captures the underlying biological nature of the bioprocess. What is often overlooked is the importance of evaluating the DDM's prediction capabilities on unseen data [16]. Unseen data is data that has not been used during the development (training, validating, testing) of the DDM. Traditionally, testing data is partitioned from the model development data and used to evaluate the model's performance on future data [17]. However, using only testing data to evaluate a model's prediction capability on future data is not sufficient for bioprocesses that face data volume and/or data variability challenges. Unseen data enables unbiased evaluation of the model's interpolation and extrapolation capabilities. This is particularly relevant if the volume of data available does not contain the degree of variability that the bioprocess may exhibit, for example, not capable of collecting data describing the feedstock subject to the external conditions described in the previous paragraph.

This research proposes a methodology for modelling bioprocesses that face data volume and variability challenges. Furthermore, the work explores how model data and predictions can be visualised to see trends and aid communication. To achieve this, two new steps are added to the traditional modelling methodology, Fig. 1. The first is to use exploratory data analysis to statistically evaluate and visualise the model data boundaries and the spread of data within the boundaries. The second is to use unseen data to evaluate how well the developed model captures the variability of the bioprocess. The proposed methodology is demonstrated by developing a DDM for an industrial case study that faces data volume and variability challenges.

## 2. Material and methods

### 2.1. The industrial $H^2$AD bioprocess

The $H^2$AD process is an industrial bioprocess plant developed by Lindhurst Engineering Ltd in collaboration with the University of Nottingham for treating wastewaters from small-medium enterprises (SME) manufacturers spanning food and drink, due to its modular, low-cost design. The $H^2$AD technology combines a bioelectrochemical system (BES) and anaerobic digestion. Bioelectrochemical systems are systems capable of converting chemical energy into electrical energy (and vice-versa) by employing microbes as catalysts [18]. Anaerobic digestion is a chain of interconnected biological reactions, where the organic matter, is transformed into methane, carbon dioxide and anaerobic biomass, in an oxygen-free environment [19].

The $H^2$AD plant studied in this work is situated at a dairy farm in the East Midlands, UK. The $H^2$AD bioprocess treats farm wastewaters containing cattle slurry, bedding waste, waste milk, footbath, parlour washings, and rainfall, which is separated by a screen press and stored in a 3000 $m^3$ slurry tank. The wastewater from the slurry tank is fed into the $H^2$AD bioprocess, contained within a shipping container, that generates bioenergy for the farm and improves the quality of that wastewater to support reuse (Fig. 2). The steps outlined in Fig. 1 were followed to develop a DDM of the system.

### 2.2. Model data

The model was developed from a set of known input and output data collected from the $H^2$AD bioprocess. The model input data originates from two sources, the $H^2$AD feedstock data and the $H^2$AD process data. To characterise the $H^2$AD feedstock, water quality analysis was performed on samples collected entering the $H^2$AD plant. A total of 17 water quality analysis (WQA) parameters were measured and are detailed in SI Table 1. In addition to this, the $H^2$AD unit automatically collected data on four $H^2$AD process conditions via sensors wirelessly connected to a database hosted on a cloud platform. These were hydraulic retention time (HRT), recirculation flow rate (F), system temperature (T) and system pressure (P). The $H^2$AD output was characterised by the percentage removal of chemical oxygen demand (COD) from the wastewater. There are 30 data-points available of WQA performed on the farm waste samples, taken from the slurry tank feeding the $H^2$AD system. In addition, there were 52 temporal data points representing one year's worth of historical $H^2$AD process data automatically collected by the unit.

To maximise the data available and demonstrate the modelling methodology developed, an additional 22 synthetic WQA data points were generated using pair-copula constructions. Combining the actual data points from the one-year historical sampling and generated

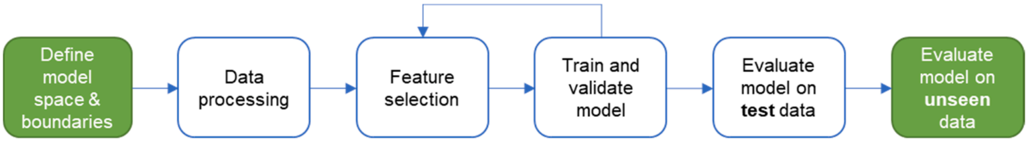
Define model space & boundaries
Data processing
Feature selection
Train and validate model
Evaluate model on test data
Evaluate model on unseen data

**Fig. 1.** The expanded bioprocess modelling methodology to include two new steps (the green boxes).

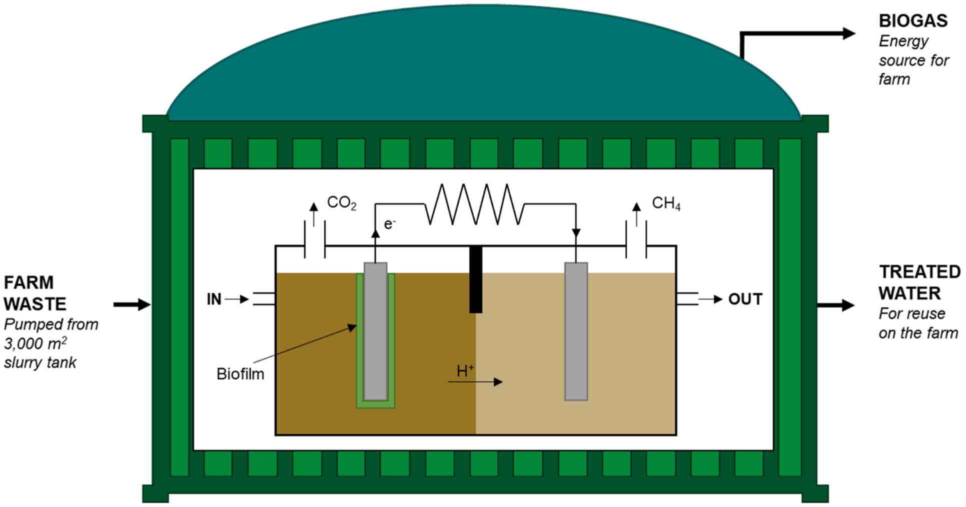

**Fig. 2.** The H$^2$AD plant and technology within it.
Adapted from [20].

synthetic data led to a total of 52 points. In situations where data from the process operation may be difficult to obtain, due to budget, time or privacy concerns, generating synthetic data poses a viable alternative [47]. When generating synthetic data the standard method is to randomly sample from a probability density function fitted to existing data [21]. Pair-copula constructions are a popular tool for generating synthetic data from multivariate distributions, due to their simple structure and high flexibility [22]. Using synthetic data has risks associated with reproducing bias inherent within the historic data and generating unrealistic data (i.e., does not capture the variability of the original data) that may reduce the generalisation capability of a DDM [22]. Therefore, it is important to validate a model built using synthetic data on additional unseen data when available. By comparing the Pearson correlation coefficients in the original dataset and synthetic dataset, shown in Fig. 3, it can be seen that the data generated widely captures the relationships in the original data.

The approach for generating synthetic output data followed that in Brissette et al., whereby trends identified from historical H$^2$AD data governed the development of non-linear equations to generate synthetic output data from the input data [23]. The dataset used to develop the DDM of the H$^2$AD system is available in the SI Table 2 and Table 3.

### 2.3. Define model space and boundaries

Exploratory data analysis was performed to understand the bioprocess represented in the model data, using a combination of statistical and data visualisation techniques. The model development data was analysed to determine the bioprocess's features: central tendency, correlations, distribution, spread, modality, and extents. The statistical measures employed were: mean, median, standard deviation, Pearson's correlation coefficient, range, maximum, and minimum. The bootstrap procedure was used to assess the variability of the model data statistics [5]. Repeated samples were taken from the model data, with replacement, to recalculate the statistic for each resample [5]. The set of sample statistics were then used to estimate the standard error. The data visualisation techniques utilised which were capable of plotting high dimensional data included: scatter plot matrix, box plots, heatmaps, and parallel coordinate plots. In order to plot variables of different magnitudes, the data was first normalised by Eq. (2), as described in Section 2.2.

### 2.4. Data processing

Data was integrated from two sources, water quality analysis performed on H$^2$AD input and output stream samples and H$^2$AD process data automatically collected remotely by the unit. Data normalisation was undertaken to ensure the impact of variables of lower magnitudes are given the same weight by the model's algorithm as the variables of larger magnitudes. The COD and total suspended solids (TSS) data have magnitudes within the order of 10,000, while for the dissolved oxygen (DO) the magnitude is much lower and in the order of 0.1. To normalise the data the minimax function was applied, as this normalises the variables without any loss of information [6]. The equation to normalise the data is given below:

$$x_{norm} = \frac{x - \quad x_{min}}{x_{max} - \quad x_{min}} \tag{2}$$

Where $x$ is the variable, $x_{norm}$ is the normalised variable and $x_{min}$ *and* $x_{max}$ is the minimum and maximum values of the variable being normalised respectively.

### 2.5. Feature selection and dimensionality reduction

To avoid the challenges that arise from modelling a high-dimensional dataset that contains a limited volume of data [6] (21 input variables, and 52 datapoints), *feature selection* and *dimensionality reduction* methods were employed. Feature selection reduces the number of variables in the dataset, keeping only the most relevant variables, whilst dimensionality reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the

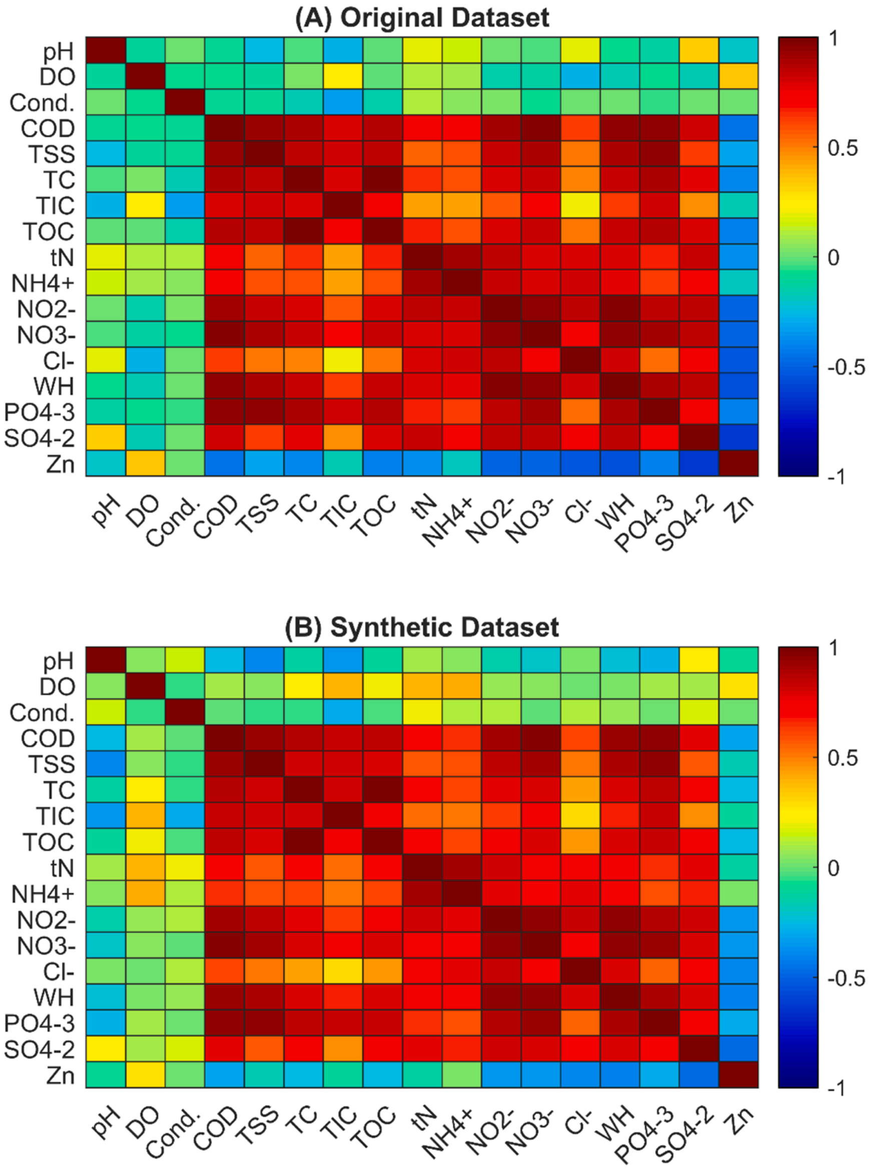

**Fig. 3.** Heatmaps displaying the correlation coefficients between variables in (A) the original water quality analysis data and (B) the original plus synthetic data [dissolved oxygen: DO; conductivity: Cond.; chemical oxygen demand: COD; total suspended solids: TSS; total carbon: TC; total inorganic carbon: TIC; total organic carbon: TOC; total nitrogen: tN; nitrite: $NO_2^-$; nitrate: $NO_3^-$; chloride: $Cl^-$; water hardness: WH; phosphate: $PO_4^{3-}$; sulphate: $SO_4^{2-}$; zinc: Zn].

low-dimensional representation retains the meaningful properties of the original data [24]. It is standard practice to perform feature selection before dimensionality reduction so that any transformations are only performed on the relevant features [24]. For this work, the feature selection technique, neighbourhood component analysis (NCA) was utilised to identify and remove the variables that had a minimum impact on the model output [25]. The minimum impact was defined as returning a feature weight of close to zero (<0.1) once fitted to the NCA model [25]. The dimensions were then further reduced by the dimensionality reduction technique, principal component analysis (PCA), that transforms potentially correlated data into an orthogonal system of linearly uncorrelated principal components (PC) [26].

The NCA for regression identified 6 input variables to remove, so that the $H^2$AD input variables are reduced from 21 to 15. The variables removed were: total inorganic carbon (TIC), total carbon (TC), nitrite ($NO2^-$), nitrate ($NO_3^-$), total nitrogen (tN) and system pressure (P). While many of these variables have a strong correlation ($r > 0.7$, Fig. 3(B)) with variables not removed by the NCA, their weak correlation ($r < 0.4$) with COD removal may explain why the NCA method identified these variables for removal (Table 1). Additionally, previous studies, summarised in SI Table 1, offer insights into the physical, chemical and/or biological phenomena as to why these variables appear to have a limited effect on the $H^2$AD performance.

The standard practice is to reduce the number of variables so that 95% of the variance within the dataset is captured by the PCs [26]. By applying PCA to the $H^2$AD dataset, the input variables were further reduced from 15 to 9 PCs.

**Table 1**
Pearson correlation between variables selected for removal by the neighbourhood component analysis and the model output, chemical oxygen removal rate [total carbon: TC; total inorganic carbon: TIC; total nitrogen: N; nitrite: NO2-; nitrate: NO3-; system pressure: P].

| TC | TIC | tN | NO2- | NO3- | P |
|---|---|---|---|---|---|
| 0.12 | -0.20 | -0.17 | -0.39 | -0.35 | -0.18 |

### 2.6. Data partitioning

The data used to develop the DDMs was then partitioned at several levels, as independent data is required to train, validate, test and evaluate the model (Fig. 4). The data is first partitioned into 80% model development data and 20% unseen data from within and outside the model's boundaries. The semi-random partitioning framework was employed to partition the unseen data points within and outside the model boundaries [27]. The model development data is further randomly partitioned into training and testing data, following the widely applied 80:20 ratio, respectively [28]. The training data was further portioned into training and validation data using the k-fold cross-validation technique [17]. Training data is the data fitted to the model's algorithm, whilst validation data provides an evaluation of the model's fit to the training data and is used for tuning the algorithm's hyperparameters. A hyperparameter is an adjustable parameter that must be either manually or automatically tuned in order to obtain a model with the optimal performance [29]. The testing data is used to evaluate the model's fit on the model development data.

### 2.7. Algorithm selection, training, validation and testing

For this work, four machine learning algorithms suitable for regression were investigated: Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural networks (ANN) (SI Table 4). These algorithms were chosen as they have been shown to typically have stronger predictive powers over other algorithms like linear regression [30,31]. Ensemble models are built from integrating the predictions made by various models into one output [32]. Ensemble modelling has shown potential for modelling bioprocesses as a way to account for uncertainty in the model structures and parameters [32]. An ensemble model was built by summing the weighted predictions of all four models according to Eq. (3). A grid search was performed to find the weight scores that produced the lowest total mean absolute error on the training and testing data.

$$p = \quad w_1p_1 + \quad w_2p_2 + \quad w_3p_3 + \quad w_4p_4 \tag{3}$$

Where $p$ is the predicted value of the ensemble model, $p_{1-4}$ is the predicted values of the individual models, and $w_{1-4}$ is the weighted score applied to each model's prediction.

Two ensemble models were built for the $H^2$AD bioprocess. The first was developed by combining the weighted predictions from the four previously developed DDMs and shall be referred to as all-ensemble (All-En) model throughout this manuscript. A grid-search was conducted to determine the combination of weights that produced the All-En model with the lowest combined MAE score on the training and testing data. The result was an All-En model that was weighted 0.1 GP model, 0.1 RF model, 0.1 SVR model and 0.7 ANN model. Due to the ANN algorithm's strong performance on the training data, a second ensemble model was developed from five individual ANN models, referred to as a stacked ANN (Stk-ANN) model. A second grid-search was performed to determine the weights of the five ANN models. The Stk-ANN model was weighted 0.2 ANN_1, 0.1 ANN_2, 0.2 ANN_3 and 0.1 ANN_4 and 0.4 ANN_5. All the ANN models had the optimal architecture determined from training the ANN model using cross-validation (one hidden layer with 7 nodes), so differences in performance is attributed to final weights between nodes after completion of the training stage.

The training, validation and testing of the DDMs and hyperparameter optimisation were carried out by MATLAB R2017.b software developed by MathWorks.

### 2.8. Metrics for statistical evaluation of models

The prediction capabilities of the various models were statistically evaluated in terms of the coefficient of determination (R2), the root of the mean squared error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). The $R^2$ is the square of the sample correlation coefficient between observed values and predicted values, and is a measure of the explained variance by the model [26]. The RMSE measures the mean square magnitude of the error [33], while MAE measures the absolute magnitude of the errors [34]. The MAPE is similar to MAE but is based on percentage error, being that it is scaled independently and can be used to compare model performance across different data sets [34]. The statistically best model is the one that has an $R^2$ closest to 1 while minimising RMSE, MAE and MAPE. The $R^2$, RMSE, MAE and MAPE are defined as follows:

$$R^2 = \quad \frac{\left(\sum_{i=1}^{n}(p_i - \overline{p}_i)(a_i - a_i)\right)^2}{\sum_{i=1}^{n}(p_i - \overline{p}_i)^2(a_i - a_i)^2} \tag{4}$$

$$RMSE = \quad \sqrt{\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{n}} \tag{5}$$

$$MAE = \quad \frac{\sum_{i=1}^{n}|a_i - p_i|}{n} \tag{6}$$

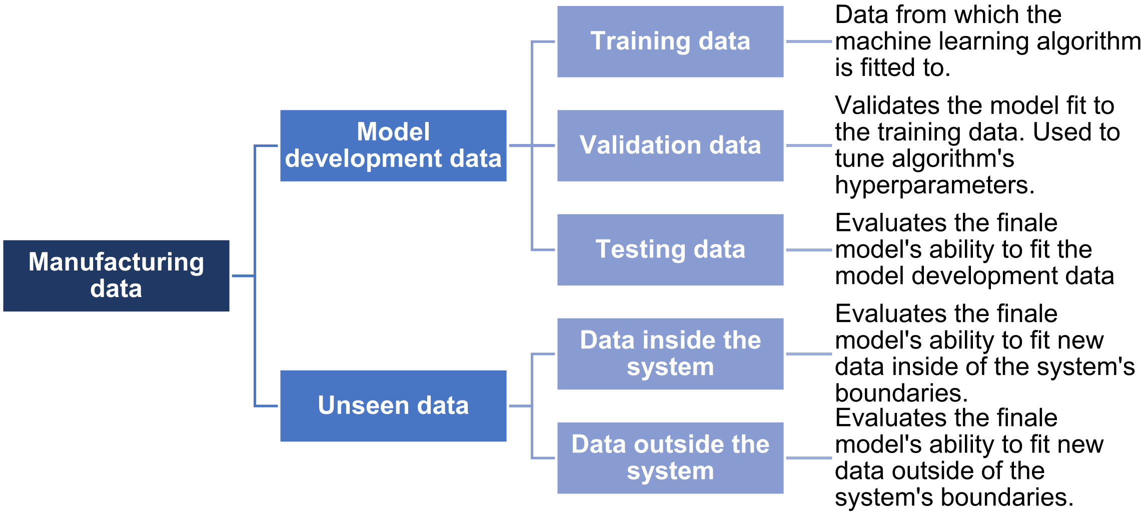
Manufacturing data
Model development data
Unseen data
Training data
Validation data
Testing data
Data inside the system
Data outside the system
Data from which the machine learning algorithm is fitted to.
Validates the model fit to the training data. Used to tune algorithm's hyperparameters.
Evaluates the finale model's ability to fit the model development data
Evaluates the finale model's ability to fit new data inside of the system's boundaries.
Evaluates the finale model's ability to fit new data outside of the system's boundaries.

**Fig. 4.** Hierarchy of data partitioning during the development of a data-driven model.

$$MAPE = 100 \quad \times \quad \frac{\sum_{i-1}^{n} \left| \frac{a_i - p_i}{a_i} \right|}{n} \tag{7}$$

Where $n$ is the number of points, $p_i$ is the predicted value, $a_i$ is the actual observed value, and the symbol $^{-}$ is the average of the related values.

## 3. Results and discussion

### 3.1. Exploring the model data space and boundaries

Data-driven models have a limited ability to extrapolate beyond the model data boundaries compared to other modelling techniques, because they learn to fit the known data as closely as possible, regardless of how it performs outside of these situations [35]. Therefore, it is important to define the space and boundaries of the model data in order to establish what portion of the bioprocess is being used, and thus how representative the driven model is of the actual process. Statistical analysis of the model data was performed to identify the data's key features, such as central tendency, variability, distribution and boundaries. As previously discussed, for a DDM to be truly representative of the bioprocess these statistics should be equal to the bioprocess parameters [5]. As the bioprocess parameters are unknown, the Bootstrap confidence intervals of the model data statistic were calculated to estimate the uncertainty that these statistics are representative of the system parameter. The results of the statistical analysis are reported in Table 1.

The central tendencies describe the typical value of a variable and provide an estimate of where most of the data is located in the system [5]. Statistically, this is achieved with either the mean or median, which have been calculated for each variable in Table 1. The variables TSS and total organic carbon (TOC) show a greater than 10% difference between the mean and median. This can indicate the mean is being influenced by extreme scores and is not an accurate representation of the data-set central point [5]. Instead, the median is more informative in representing the centre of the model space. The bootstrap 95% confidence intervals (CI) have been calculated for the mean and median score of each variable in Table 1. The CI is used to assess the variability of each statistic. Statistics that contain greater variability would imply greater uncertainty that the model data is representative of the system for that variable [5]. For example, the pH mean and median scores have the smallest percentage CI (mean: −0.53% and 0.67%; median: −0.67% and 0.40%), thus increasing the confidence that the pH statistics are representative of the system parameters. Out of all the model input data, only the pH, conductivity, HRT, PS and T have confidence intervals around their mean and median statistics less than 5%. This illustrates the high level of variability in the bioprocess and indicates that more data may be needed to trust conclusions drawn from the model.

As previously stated, when modelling bioprocesses there is a tendency to overlook defining the model data space or boundaries [6,26,36,37] and those that do typically make little attempt to quantify any uncertainty regarding these values [38–40]. Defining the central tendencies and their uncertainties for the model data informs what values we expect to observe in new data points collected. If the central tendencies of new data are statistically significantly different (determined via hypothesis testing), there may have either been a change to the bioprocess or the model data was never fully representative of the bioprocess. For both situations, the model is not representative of the bioprocess and requires retraining.

The variability and spread of model data around the variables' mean, is described by the standard deviation [5]. The relative standard deviations (the percentage spread around the mean) for each variable has been calculated and reported in Table 1. The pH and conductivity have smaller relative standard deviations than the other feedstock variables (2.20 and 5.59 compared to 17.6 +), implying they are more closely clustered around the mean and contain less variability. Whereas, the remaining feedstock variables have higher relative standard deviations, indicating that the data is more spread out [5]. Data visualisation techniques, such as a scatter plot matrix and parallel coordinate plot, have been utilised alongside the statistics to explore the data and visualise the spread of data throughout the model boundaries.

#### 3.1.1. Visualising the model boundaries

Data visualisations can be more effective than words or numbers at conveying key information, provided the visualisations are an honest representation of the data [41]. The visuals produced can aid understanding of what portion of the bioprocess the model represents. Scatterplot matrices are utilised to visualise a manufacturer's data to identify potential trends, correlations, and clusters [42,43]. A scatter plot matrix was created to visualise the $H^2AD$ bioprocess data in Fig. 5. The histograms within Fig. 5 indicates the TOC, ammonium ($NH_4^+$), HRT, PS, T and %COD removal variables contain regions that are sparsely populated. This is especially true for the HRT variable where more than 90% of the data is between the range of 3.5 and 4 days and only two data points are collected when the $H^2AD$ system is operating at more than 6 days. Similarly, the PS variable was operated at 4 distinct speeds with 2% of that data recorded at 8.95 L/min compared to more than 60% of the data recorded at 9.94 L/min. The scatter plot matrix is a useful tool for communicating regions that are underrepresented by the model. Scatterplot matrices are also used to identify potential trends in multivariate data [44]. For example, Fig. 5 indicates positive intercorrelations between the variables TSS, COD, TOC, water hardness (WH), phosphate ($PO_4^{-3}$) and sulphate ($SO_4^{-2}$). These relationships can then be quantified by calculating the Pearson correlation coefficient between the variables.

Visualising a DDM boundaries will inform process engineers whether new data points fit within or outside model boundaries, helping to build confidence in a model's predictions. A useful tool for visualising all the systems variables in one figure is the parallel coordinate plot (PCP). The PCP is able to visualise high dimensional data by, setting each variable along a point on the x-axis [45]. Data-points are normalised, so all values can be displayed on a single y-axis, and plotted as a series of lines, displaying a data-point normalised value for each variable [45]. By adapting the PCP, a traffic light classification was developed to quickly warn if variables of a new sample are within the boundaries of the model development data (green region), approaching the model boundaries (yellow region) or outside the model boundaries (red region), see Fig. 6.

The red boundaries are defined as the maximum and minimum of the model development data and the yellow boundaries the 0.2 and 0.8 quantiles. In Fig. 6, two examples from the data are given for a data-point that remained in the model boundaries, and a data-point that was outside of the model boundaries, for certain variables. The advantage of this traffic light classification is that it provides a simple visualisation that can be displayed on a handheld device and provide the operators with real-time information that the system is currently operating within the boundaries that the model was developed on. Therefore, informing them which model predictions can be trusted.

While the red boundaries will always be the maximum and minimum value of the model's input variables, the yellow boundaries may vary depending on the requirements of the model predictions and how accurate the model is at making predictions outside of the model boundaries. For example, when the GP model was tested on unseen data outside of the model boundaries (Section 3.3), it performed relatively well (MAE 5.58) and when predicting the percentage of COD removal, it meets the requirement. However, if this model was built for predicting an output where a greater level of accuracy is required (e.g. predicting a safety-critical variable), the yellow boundaries may have to be increased to give an earlier warning to the operators that the bioprocess is approaching the point outside the boundaries of the data the model was trained on.

### 3.2. Analysis of the data-driven models' results

With respect to predicting the $H^2AD$ bioprocess to remove COD from

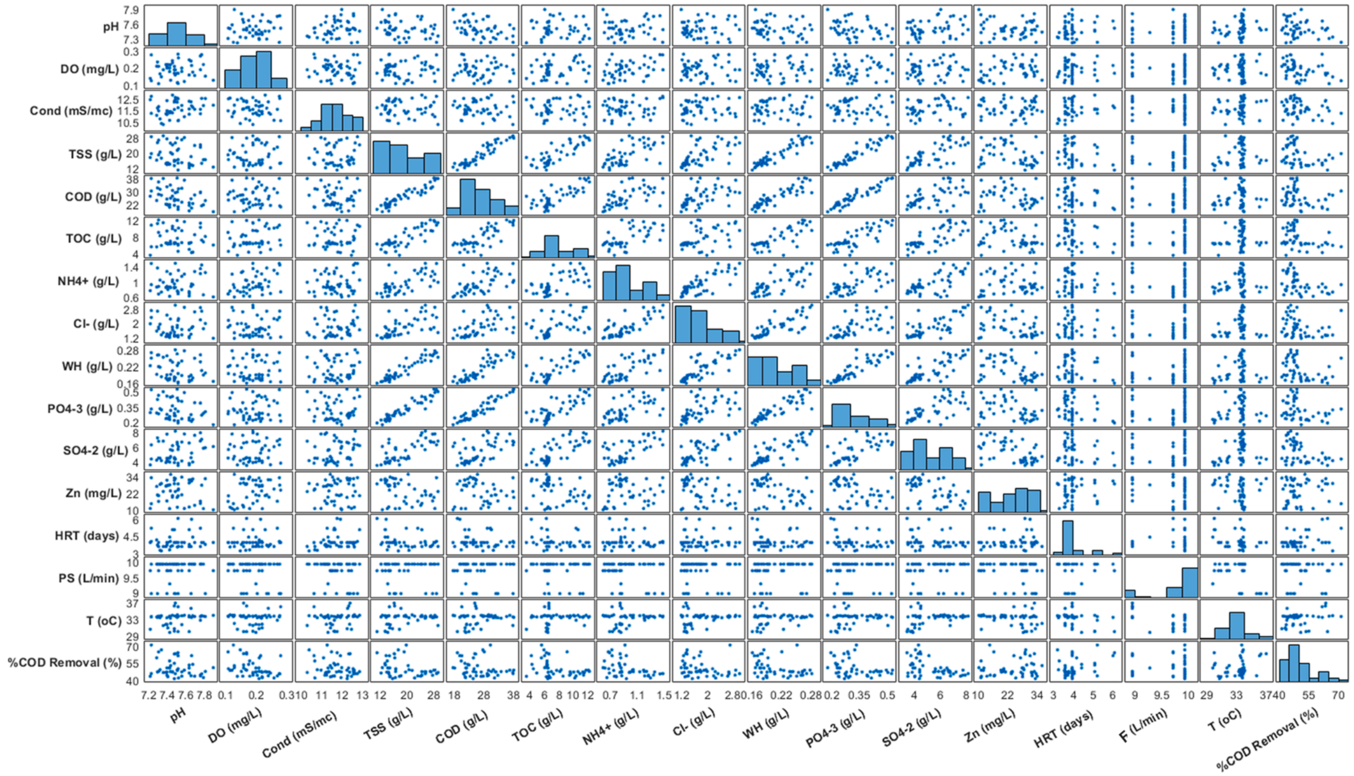

**Fig. 5.** Scatter plot matrix of the model data [dissolved oxygen: DO; conductivity: Cond.; chemical oxygen demand: COD; total suspended solids: TSS; total organic carbon: TOC; ammonium: NH$_4^+$; chloride: Cl$^-$; water hardness: WH; phosphate: PO$_4^{3-}$; sulphate: SO$_4^{2-}$; zinc: Zn, hydraulic retention time: HRT; recirculation flow rate: F; system temperature: T; chemical oxygen removal rate: %COD Removal].

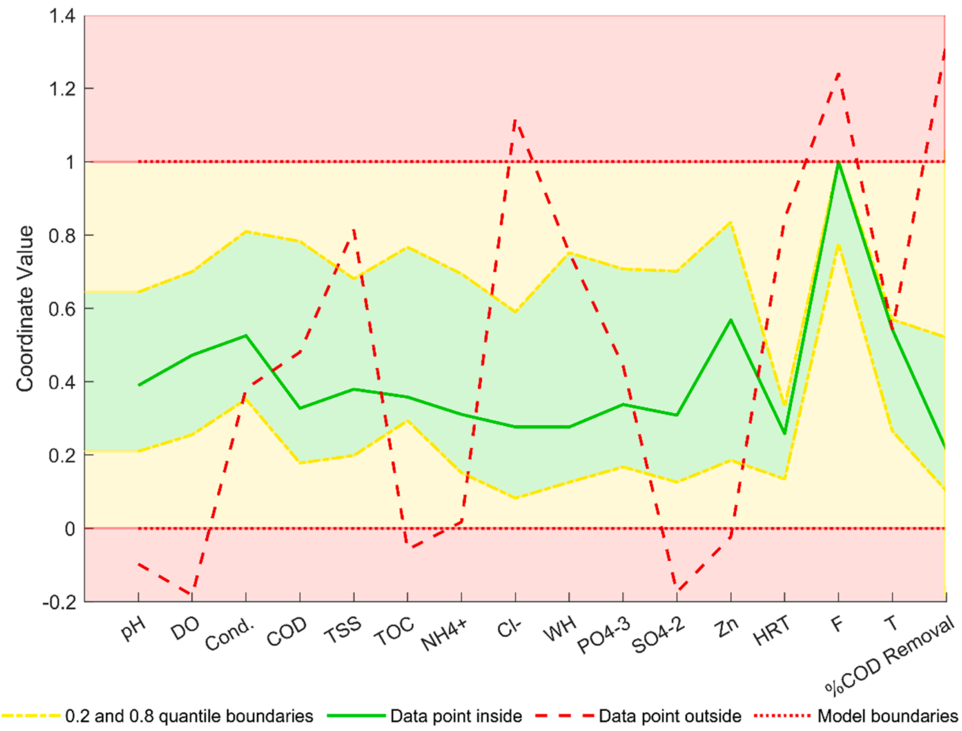

**Fig. 6.** Traffic light system warning if new data points are within the model boundaries (green region), approaching the model boundaries (yellow region) or outside the boundaries (red system) [dissolved oxygen: DO; conductivity: Cond.; chemical oxygen demand: COD; total suspended solids: TSS; total organic carbon: TOC; ammonium: NH$_4^+$; chloride: Cl$^-$; water hardness: WH; phosphate: PO$_4^{3-}$; sulphate: SO$_4^{2-}$; zinc: Zn, hydraulic retention time: HRT; recirculation flow rate: F; system temperature: T; chemical oxygen removal rate: %COD].

wastewaters, several algorithms were evaluated for their ability to fit the bioprocess data, including Gaussian process (GP), random forest (RF), support vector regression (SVR) and artificial neural networks (ANN). The model development data (Fig. 4) was used to determine the optimal hyperparameters and model parameters of the DDMs. The 42 model development data points were randomised and 80% (34 data points) were partitioned for the training and validation of the model and the remaining 20% of data (8 data points) were partitioned for testing the final model. A Bayesian optimisation was performed to optimise the hyperparameters, using k-fold cross-validation to evaluate each configuration's ability to fit the validation data. For the k-fold cross-validation, a k value of 10 was chosen because of the limited number of data points [46]. The models' ability to fit new data was then evaluated using the testing data points. The results from the statistical analysis of the models' ability to fit the model training and testing data are reported in Table 2.

It has been common practice within process engineering to use regression plots, of actual values against a model's predictions when determining how well the model fits the data [26]. Regression plots are useful for a quick guide in determining how accurately a model's predictions match the actual data. Inspection of the regression plots' gradients in Fig. 7 indicates that the GP, SVR and ANN algorithms achieved a closer agreement (gradients $> 0.81$) between the actual and predicted training data points compared to the RF algorithm (gradient $= 0.28$). This conclusion is also supported by the statistical analysis in Table 2, as the RF model has the largest MAE and MAPE values (3.84 and 7.39 respectively) and the second smallest $R^2$ value (0.79). The exception is the RMSE error metric, of which the RF has the second smallest value (1.03). This may be explained by the fact that the RMSE squares the error magnitude meaning outliers have a greater influence on the RMSE than on error metrics [34]. Random forest models typically have lower RMSE values than other algorithms because an RF prediction is the average of hundreds of decision tree models; therefore, reducing the occurrence of outliers [47]. The RF model's MAPE and MAE scores are inconsistent with the evidence in the literature that show RF models to have similar performance to ANN and SVR models when modelling BES [48,49]. To understand this observation, residual plots were produced to explore how the models perform throughout the bioprocess. The residuals are calculated as the actual value minus the predicted percentage removal of COD, and the plot for each model are presented in Fig. 7. The residual should be randomly dispersed around the horizontal axis for the model to be considered a good fit to the data [47]. The residual plot of the RF model (Fig. 7 B2) displays a positive linear relationship between the residuals and actual values, indicating there are underlying faults in the RF model. The statistical analysis shows the GP algorithm has the closest fit to the training data ($R^2$: 0.99; RMSE: 0.01; MAE: 0.01; MAPE: 0.01). On inspection of the GP residual plot (Fig. 7 A2), the model appears to have overfitted the training data. However, this is a characteristic of GP models due to the algorithm posterior probability fit [50]. It is therefore vital to evaluate these models on testing data to get an indication of the model's ability to describe the bioprocess. The residuals in the SVR and ANN residual plots (Fig. 7 C2 and Fig. 7 D2 respectively) appear to be randomly dispersed around the horizontal axis for both models, suggesting they do not contain the faults and are homoscedastic. These results concur with previous evidence, which indicates that SVR and ANN to be suitable algorithms to fit BES data [49, 51]. However, further evaluation is required by using the testing data to investigate whether these patterns are restricted only to the training data.

Evaluating the model on testing data is necessary to determine if the model has overfitted on the training data [17] and to assess the model's ability to predict future data. The GP algorithm had the best statistical performance on the training data, but the residual plot (Fig. 7 A2) indicated that the model may have overfitted on the training data. This conclusion is supported by the drop in performance from fitting the training data to fitting the testing data ($R^2$: 0.99 → 0.80; RMSE: 0.01 → 3.15; MAE: 0.01 → 4.88; MAPE: 0.01 → 8.49). The ANN algorithm shows the best performance at fitting the testing data, dominating three of the four error metrics ($R^2$: 0.93; RMSE: 1.77; MAE: 3.64; MAPE: 4.17). These results, coupled with the ANN algorithm's strong performance on the training data, proves that the ANN is best suited to modelling the H$^2$AD bioprocess. However, the ANN accuracy ($R^2$: 0.93) on the testing

**Table 2**
The model data statistics and their bootstrap 95% confidence intervals in square brackets [dissolved oxygen: DO; conductivity: Cond.; chemical oxygen demand: COD; total suspended solids: TSS; total organic carbon: TOC; ammonium: $NH_4^+$; chloride: $Cl^-$; water hardness: WH; phosphate: $PO_4^{-3}$; sulphate: $SO_4^{-2}$; zinc: Zn, hydraulic retention time: HRT; recirculation flow rate speed: F; system temperature: T; chemical oxygen removal rate: %COD Removal].

| | Mean | Median | Min | Max | Relative Standard deviation |
|---|---|---|---|---|---|
| pH | 7.5 | 7.49 | 7.22 | 7.9 | 2.20 |
| | [7.46 7.55] | [7.44 7.52] | [7.22 7.23] | [7.81 7.90] | [1.87 2.63] |
| DO | 0.192 | 0.190 | 0.110 | 0.28 | 24.0 |
| (mg/L) | [0.178 0.204] | [0.170 0.210] | [0.110 0.111] | [0.277 2.80] | [20.4 27.9] |
| Cond. (mS/mc) | 11.7 | 11.6 | 10.3 | 12.8 | 5.89 |
| | [11.5 11.9] | [11.4 11.9] | [10.3 10.4] | [12.7 12.8] | [4.75 6.55] |
| TSS | 19,100 | 17,000 | 11,100 | 29,200 | 27.9 |
| (mg/L) | [17,600 20,500] | [15,700 20,100] | [11,100 12,300] | [28,700 29,200] | [24.3 31.8] |
| COD | 26,600 | 25,300 | 17,400 | 38,200 | 20.8 |
| (mg/L) | [25,000 28,200] | [24,300 27,900] | [17,400 18,000] | [37,900 38,200] | [17.8 24.9] |
| TOC | 7680 | 6640 | 3540 | 12,200 | 30.1 |
| (mg/L) | [7060 8320] | [6480 7920] | [3540 4220] | [12,100 12,200] | [25.4 34.8] |
| $NH_4^+$ | 960 | 880 | 604 | 1490 | 25.8 |
| (mg/L) | [899 1040] | [854 992] | [604 623] | [1490 1490] | [22.0 30.2] |
| Cl- | 1810 | 1690 | 1190 | 3020 | 29.1 |
| (mg/L) | [1680 1980] | [1500 1850] | [1190 1220] | [2900 3020] | [24.7 35.2] |
| WH | 209 | 194 | 160 | 281 | 17.6 |
| (mg/L) | [199 220] | [184 222] | [160 165] | [277 281] | [15.6 20.1] |
| $PO_4^{-3}$ | 326 | 297 | 179 | 529 | 29.6 |
| (mg/L) | [298 352] | [273 335] | [179 197] | [524 529] | [25.4 34.7] |
| $SO_4^{-2}$ | 5360 | 4920 | 3480 | 8170 | 26.1 |
| (mg/L) | [4980 55,790] | [4440 5960] | [3480 3540] | [7940 8170] | [23.1 30.2] |
| Zn | 23.7 | 24.8 | 10.3 | 35.8 | 32.6 |
| (mg/L) | [21.4 25.6] | [20.4 28.2] | [10.3 11.1] | [33.8 35.8] | [29.2 37.8] |
| HRT | 4.05 | 3.95 | 3.21 | 6.09 | 15.7 |
| (days) | [3.90 4.24] | [3.87 3.95] | [3.21 3.26] | [6.02 6.09] | [11.4 21.1] |
| F | 9.72 | 9.94 | 8.95 | 9.94 | 3.78 |
| (L/min) | [9.59 9.80] | [9.72 9.94] | [8.95 8.95] | [9.94 9.94] | [2.91 4.56] |
| T | 33.2 | 33.7 | 29.8 | 37.0 | 4.85 |
| (oC) | [32.7 33.6] | [33.6 33.8] | [29.8 30] | [36.3 37.0] | [4.01 6.05] |
| % COD removal | 50.6 | 47.7 | 41.2 | 71.2 | 14.8 |
| (%) | [48.8 52.9] | [47.1 50.7] | [41.2 41.9] | [66.9 71.2] | [12.2 18.2] |

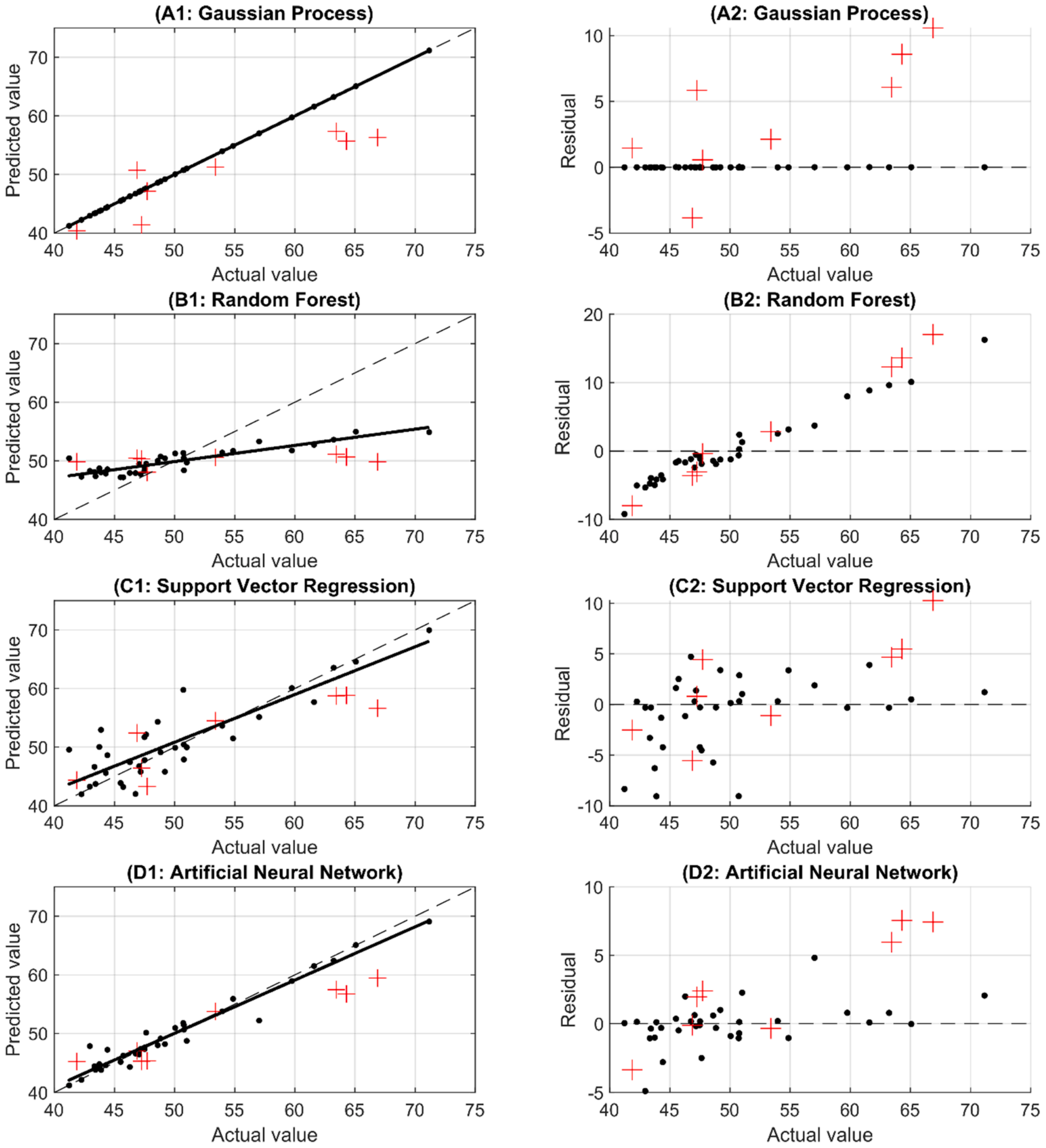

**Fig. 7.** The predicted percentage of COD removal plotted against the actual value (1) and the residuals plotted against the actual values (2), for (A) Gaussian process, (B) random forest, (C) support vector regression and (D) artificial neural network data-driven models. The black dots represent the training data and the red crosses the testing data.

data is lower than some ANN models in the literature ($R^2$: 0.96 [52], $R^2$: 0.95 [53] and $R^2$: 0.99 [54]). The higher accuracy reported by these models is likely caused by the models being developed from data that was collected under laboratory conditions, therefore, not subject to the same degree of variability exhibited by the data collected from the $H^2$AD plant. The $H^2$AD plant is subject to additional variability caused by external environmental conditions (local agronomic conditions, seasonal, and changes to farm practices). Ensemble models were built in an attempt to improve the accuracy of the $H^2$AD DDM.

### 3.2.1. Analysis of ensemble data-driven models results

To explore whether ensemble modelling techniques may yield more accurate predictions, two ensemble models of the $H^2$AD bioprocess were built. Comparing the ensemble model to the single ANN model, which was previously determined to have best fitted the model data, both models have similar $R^2$ scores on the training data (All-En: 0.97; ANN: 0.95) and testing data (All-En: 0.92; ANN: 0.95). The All-En model has lower RSME values (training: 1.16; testing: 1.65) than the ANN model (training: 1.48; testing: 1.77). However, the ANN model outperforms the All-En model on the MAE and MAPE scores for both the training (ANN_MAE: 1.00; ANN_MAPE: 2.03; ALL-En_MAE: 1.18; All-En_MAPE: 2.36) and testing data (ANN_MAE: 3.64; ANN_MAPE: 6.30; ALL-En_MAE: 4.02; All-En_MAPE: 6.92), strengthening the argument that models developed using ANN are better suited to modelling the $H^2$AD bioprocess when compared to SVR, RF and GP algorithms. The results in Table 3 show that the Stk-ANN model outperforms both the All-En model and the ANN model for both the training and testing data. The statistical analysis is validated by graphical analysis of the regression plots in Fig. 8. The plots show that the Stk-ANN regression line has a gradient of 0.92, which is larger than the All-En model's gradient of 0.81. The residuals in both models' residual plots (Fig. 8 A2 and Fig. 8 B2) appear to be randomly dispersed around the horizontal axis for both models, suggesting they do not contain any underlying faults.

When evaluated on the testing data, the Stk-ANN model displays a

**Table 3**
Statistical analysis of the models' ability to fit the training and testing data [Gaussian process: GP; random forest: RF; support vector regression: SVR; artificial neural network: ANN; combined ensemble model: All-EN; stacked artificial neural network: Stk ANN; coefficient of determination: $R^2$; root mean squared error: RMSE; mean average error: MAE; mean absolute percentage error: MAPE].

| | $R^2$ | | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | Training data | Testing data | Training data | Testing data | Training data | Testing data | Training data | Testing data |
| GP | 0.99 | 0.80 | 0.01 | 3.15 | 0.01 | 4.88 | 0.01 | 8.49 |
| RF | 0.79 | 0.14 | 1.03 | 0.93 | 3.84 | 7.60 | 7.39 | 13.17 |
| SVR | 0.75 | 0.76 | 3.44 | 3.36 | 2.61 | 4.35 | 5.51 | 7.76 |
| ANN | 0.95 | 0.93 | 1.48 | 1.77 | 1.00 | 3.64 | 2.03 | 6.30 |
| All-EN | 0.97 | 0.92 | 1.16 | 1.65 | 1.18 | 4.02 | 2.36 | 6.92 |
| Stk ANN | 0.98 | 0.98 | 1.12 | 1.29 | 0.83 | 2.27 | 1.70 | 4.08 |

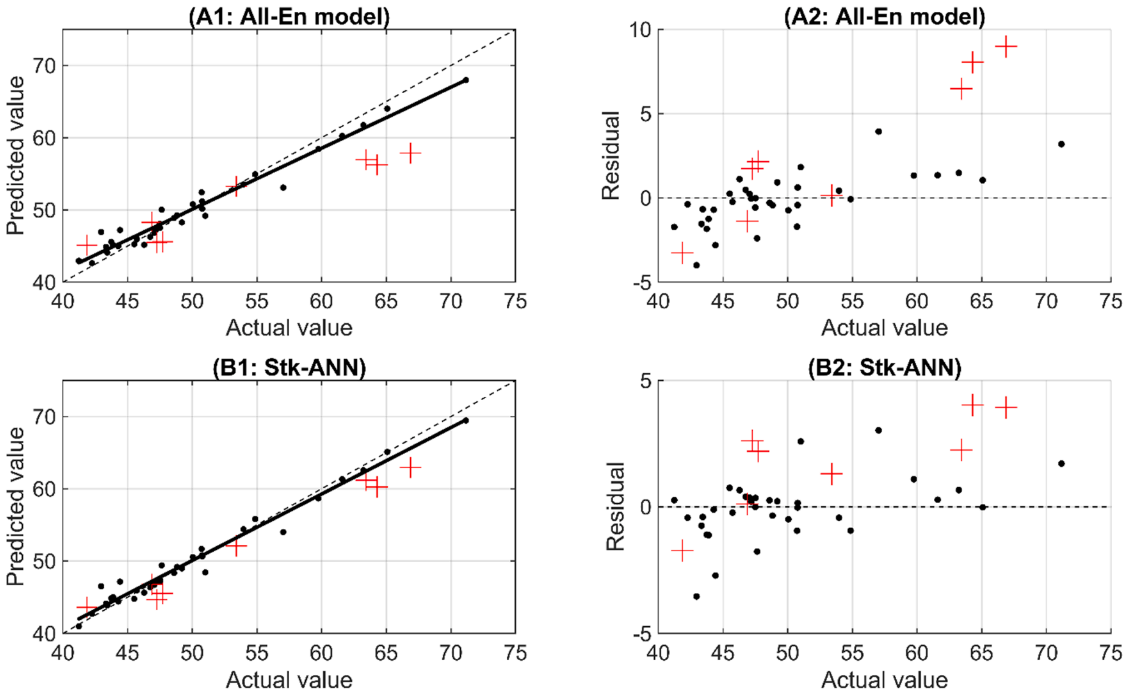

**Fig. 8.** The predicted percentage of COD removal plotted against the actual value (1) and the residuals plotted against the actual values (2), for combined ensemble model (All-En model) and (B) stacked artificial neural network models (Skt-ANN). The black dots represent the training data and the red crosses the testing data.

similar performance ($R^2$: 0.98; RMSE: 1.29; MAE: 2.27; MAPE: 4.08) to other DDMs built to study a BES [52,55]. Lesnik and Liu built an ANN model predicting the COD removal by a microbial fuel cell and when tested on testing data achieved MAPE value of 4.07, similar to the Stk-ANN model built for this work which achieved a MAPE of 4.08 [55]. Garg et al. built and evaluated three DDMs of a BES; utilising GP, SVR and ANN machine learning algorithms [52]. After evaluating the DDMs on testing data, Garg et al. observed the GP model to have the best overall performance [52]. Garg et al.'s GP model had a similar performance of the testing data ($R^2$: 0.98) as the Stk-ANN developed for this work ($R^2$: 0.98). What neither of these studies did is evaluate the DDMs' prediction capabilities on unseen data. Without evaluating on unseen data, it is harder to assess the prediction capabilities of a DDM on new data and whether the DMM has captured the variability of the bioprocess. This is vital if the model's predictions are to be trusted by a manufacturer.

### *3.3. Evaluating the models' predictive capability on unseen data*

Applying George Box's aphorism "all models are wrong, but some are useful" to manufacturing, helps to illustrate that while it is not possible to perfectly model the manufacturing system, a sufficiently representative model may aid in decision-making, develop the scientific understanding about the system, communicate knowledge and/or make predictions [56]. However, the challenge is deciding which models are useful and which are not, to avoid giving the manufacturer false or misleading information that could adversely impact on process performance. Part of the solution is to evaluate how well the model performs on new unseen data. Ten unseen data points were partitioned from the available data set and so not used in model development to evaluate the predictive capabilities of the DDMs. These unseen data points are further partitioned into points inside and outside of the model boundaries to give a comprehensive assessment of the models' predictive capabilities across and beyond the model space. The ability of the model to predict the percentage of COD removal in wastewater was statistically evaluated using, $R^2$, RMSE, MAE and MAPE and the results are reported in Table 4.

Statistical analysis of the models' ability to fit unseen data *inside* the model boundaries, further evaluates the model ability to capture the variability of the system, within the model boundaries. In Fig. 9, the models' ability to fit the testing data and unseen data inside the model boundaries have been plotted to visualise any decrease in performance when predicting new values. The Fig. 9 highlights a drop in performance as measured by the $R^2$ and RMSE error metrics for all models, excluding the SVR model. The SVR model's $R^2$ and RMSE values instead improve from 0.76 and 3.36–0.91 and 2.43, respectively, which is contradictory to previous models of process manufacturing systems that show SVR

**Table 4**
Statistical analysis of the models' ability to unseen data inside and outside the model data boundaries [Gaussian process: GP; random forest: RF; support vector regression: SVR; artificial neural network: ANN; ensemble model: En; stacked artificial neural network: Stk ANN; coefficient of determination: $R^2$; root mean squared error: RMSE; mean average error: MAE; mean absolute percentage error: MAPE].

| | $R^2$ | | RMSE | | MAE | | MAPE | |
|---|---|---|---|---|---|---|---|---|
| | Unseen data inside | Unseen data outside | Unseen data inside | Unseen data outside | Unseen data inside | Unseen data outside | Unseen data inside | Unseen data outside |
| GP | 0.75 | 0.95 | 4.25 | 5.73 | 4.22 | 5.58 | 8.52 | 9.55 |
| RF | 0.16 | 0.72 | 1.87 | 1.20 | 4.65 | 21.53 | 9.11 | 32.60 |
| SVR | 0.91 | 0.99 | 2.43 | 2.61 | 3.96 | 6.04 | 7.97 | 8.51 |
| ANN | 0.59 | 0.90 | 4.28 | 7.82 | 3.24 | 8.77 | 6.82 | 13.12 |
| En | 0.77 | 0.93 | 2.87 | 6.22 | 2.38 | 8.91 | 5.11 | 12.76 |
| Stk ANN | 0.82 | 0.94 | 2.57 | 5.74 | 1.75 | 8.08 | 3.68 | 11.69 |

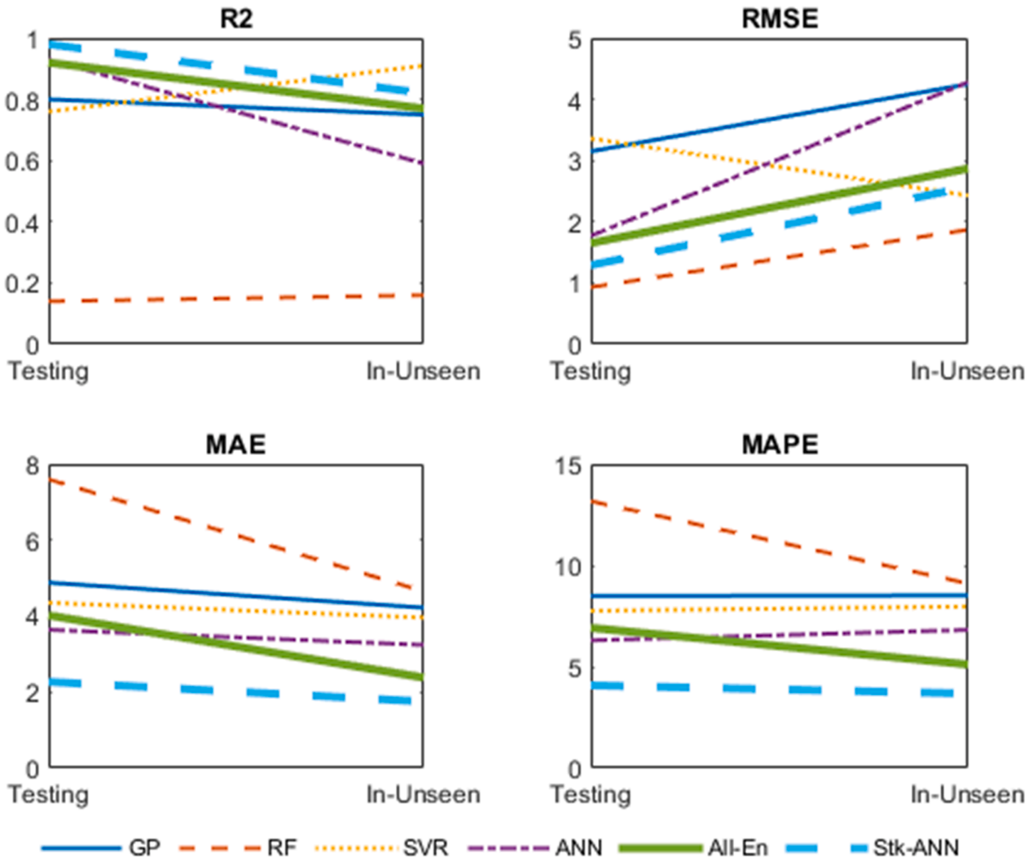

**Fig. 9.** Statistical analysis of the data-driven models to fit the testing data and unseen data inside the model boundaries (In-Unseen) [Gaussian process: GP; random forest: RF; support vector regression: SVR; artificial neural network: ANN; ensemble model: En; stacked artificial neural network: Stk ANN; coefficient of determination: $R^2$; root mean squared error: RMSE; mean average error: MAE; mean absolute percentage error: MAPE].

models perform worst on unseen data [57]. However, the drop in performance is mitigated by the fact all models demonstrate a similar performance between datasets for the MAE and MAPE values, as shown in Fig. 9. The Stk-ANN model was identified as the best model at predicting the COD removal by the $H^2AD$ bioprocess, as defined by the MAE and MAPE values for both datasets (Fig. 9). Furthermore, the Stk-ANN model is consistently one of the best performing as defined by the $R^2$ and RMSE values for both datasets. This conclusion supports prior work that suggests ANN based models of process manufacturing have a strong generalising capability when directly compared to other algorithms [58, 59].

The capability of the model to predict unseen data *outside* of the system was statistically analysed to evaluate the models' ability to extrapolate beyond boundaries of the model data. This is necessary as process manufacturing systems are continuously evolving (e.g. changes in feedstock, suppliers, equipment, operating conditions); therefore, it is necessary to understand the model capabilities to adapt to shifts in the system operating region [8]. Data-driven models have previously been shown to have poor extrapolating abilities compared to other modelling techniques [35]. However, by evaluating their exploiting ability on unseen data, it possible to quantify any decrease in performance, informing the process manufacturer to what degree they can trust the model's performance outside the model boundaries.

The $R^2$ metric does not provide a clear reflection of the models' ability to predict outside of the system, as the metric appears to have been misled by a chance correlation between the residuals that has inflated the $R^2$ score. The evidence for this statement lies in the RMSE, MAE and MAPE statistical analysis reported in Table 4 showing a decrease in performance for all the models when predicting using unseen data outside the model boundaries compared to unseen data inside the model boundaries. Whereas, the $R^2$ values that have improved for all the models. This illustrates the risk in using the $R^2$ to evaluate a model's predictive capability, the $R^2$ is susceptible to produce misleading results when modelling non-linear systems [60]. Relying on only $R^2$ to assess a model developed of a non-linear system will result in selecting the true model only 28–43% of the time [61]. This highlights the importance of using multiple statistical evaluation metrics, despite examples that relies overly on $R^2$ for evaluation [26].

The Stk-ANN produced the best results when making predictions on unseen data inside the model boundaries. However, the Stk-ANN performance when predicting data outside of the model boundaries (RMSE: 5.74; MAE: 8.08; MAPE: 11.69) is outstripped by both the GP (RMSE: 5.73; MAE: 5.58; MAPE: 9.55) and SVR (RMSE: 2.61; MAE: 6.04; MAPE: 8.51) models. In order to interpret these results, the residuals plots for the models predicting on unseen data have been produced in Fig. 10. The GP and SVR models' predictions have smaller residuals (GP: 3.58 and 9.88; SVR: 9.26 and 10.85) between the predicted and actual values for COD removal rates *larger* than the model boundaries compared to the Stk-ANN model (8.56 and 17.61). Whereas, all three models have similar residuals between the predicted and actual values for COD removal rates *smaller* than the model boundaries (GP: 2.11 and −6.74; SVR: 1.13 and −2.91; Stk-ANN: 2.67 and −3.47). For practical deployment of the model, it would therefore be appropriate to use the GP or SVR model when making predictions outside of the model boundaries and use the Skt-ANN when making predictions inside the model boundaries. This aligns with methods that develop machine learning models that are "experts" for different subsections of the system space [62]. For the $H^2$AD process manufacturing system, it is more important to be accurate when predicting the lower COD removal rates than the high COD removal rates. When treating process manufacturing wastewaters, pollutants must be removed to below a certain limit to avoid environmental damage and potential fines [63]. Overestimating the COD removal rates by the $H^2$AD bioprocess, when the actual removal rates are low, may result in pollutants not being removed to below the required limits. Therefore, the Stk-ANN limited ability to predict COD removal rates greater than the model boundaries (71%) should not be a major concern for the process manufacturer. There are limited examples of using unseen data to evaluate process manufacturing models' ability to extrapolate. One example modelled the capability of biochars for the remediation of heavy metals from water [64]. Random forest and ANN models were trained to predict the removal of five heavy metals from wastewater, before evaluating the models' ability to predict the removal

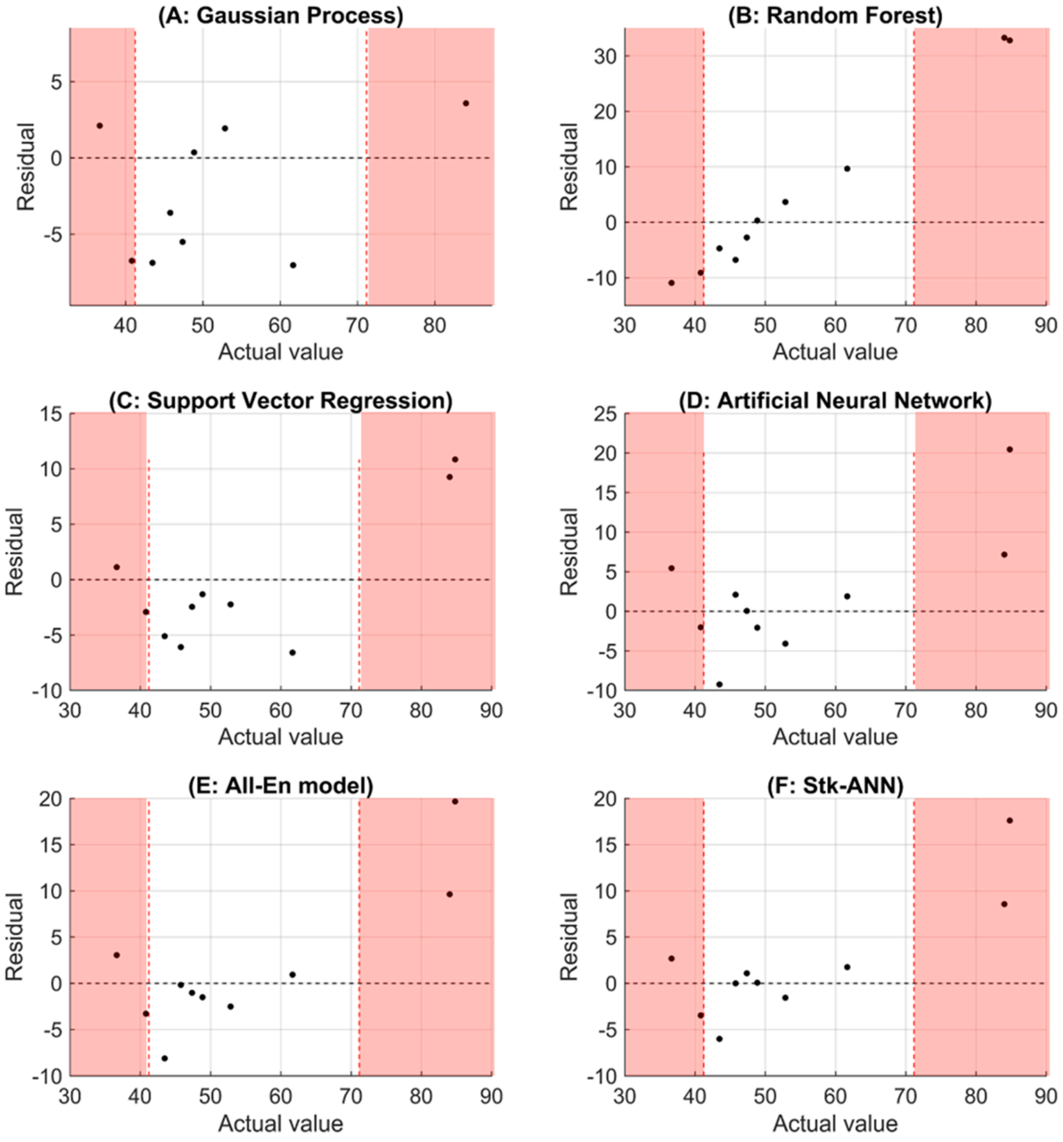

**Fig. 10.** Residuals of unseen predictions plotted against actual values for (A) Gaussian process, (B) random forest, (C) support vector regression, (D) artificial neural network, (E) combined ensemble model (All-En model) and (F) stacked artificial neural network (Skt-ANN). The red region is outside of the model data boundaries.

of an unseen metal [64]. The RMSE scores for the RF and ANN models decreased by 4% and 34%, respectively, when evaluated on the unseen metal [64]. Whereas the RMSE of the Stk-ANN model produced in this work decreased by 345%, indicating a worst extrapolating capability. This suggested that the model predictions should not be trusted when predicting outside the model data boundaries and the process manufacturer should use the parallel coordinate plot tool presented in Fig. 6 to evaluate whether future data falls outside of the model boundaries.

Fig. 10 suggests a positive correlation exists between the distance of unseen data from outside model data boundaries and the residual magnitude. This observation was validated by the very strong positive correlation between distance and residual for all the models ($0.84 < r < 0.99$). This concurs with previous research that demonstrates how the predictive performance of the data-driven models deteriorates as the system move beyond the boundaries of the data used to the model [35]. This further validates the need for Fig. 6 to assess whether new data is within or outside of a model's boundary. Furthermore, it would be possible to combine this information with Fig. 6 to provide an estimate of the prediction confidence for new data points outside of the model boundaries.

## 4. Conclusions

The development of data-driven models (DDM) for bioprocesses plays a significant role in understanding relationships within the process and making predictions. However, by using a data-driven approach, instead of for example a first principal modelling or generalised analytical modelling approach, this can limit applicability of the models to the data they were used to develop from. For a DDM to be trusted, however, the model's predictive capabilities must be evaluated. This is of greater importance to bioprocesses where either (1) data is limited because models built from limited data are in danger of overfitting to the model development data and/or (2) where variability is exhibited especially in feedstock(s). This will become more relevant as society transitions towards a circular economy where more waste streams will be used as a feedstock and come with greater variability in characteristics and composition. To avoid overfitting, the developed DDM's prediction capabilities should be evaluated on unseen data, taken both inside and outside of the model boundaries. The purpose of this work was to add two additional steps to the DDM methodology to include defining the model boundaries and evaluating the model on unseen data within and outside these boundaries. These changes are aimed at increasing a manufacturer's confidence that the model accurately represents the bioprocess being modelled. These changes will have two key benefits:

(1) **Determining the DDM's interpolation and extrapolation capabilities**: knowing the model boundaries allows unseen data to be sampled from inside and outside the bioprocess. This can then be used to assess the model's interpolation and extrapolation capabilities.

(2) **Knowing when to retrain the model**: Manufacturing systems are continuously evolving and a DDM built from one set of data will likely deteriorate with time, as the bioprocess moves beyond the original defined bioprocess. By defining the model boundaries, manufacturers can utilise visualisation techniques, like PCP, to monitor when the system moves beyond the model boundaries. The model will then need retraining, incorporating the new data into the model development data. A traffic light protocol for easily classifying if new data collected is inside or outside the model boundaries was developed for this work.

In order to demonstrate and evaluate the proposed framework, a DDM was developed of an industrial bioprocess. This study compares the performance of four machine learning algorithms, GP, RF, SVR and ANN. The algorithms' capability to predict the percentage of COD removal from wastewater by the $H^2AD$ bioprocess was evaluated. It was found that all the models performed well at fitting the model data, with the ANN model performing the best. Two ensemble models were then built to try and increase the DDM's predictive capability. Of these the Stk-ANN had the best overall performance on the training data ($R^2$: 0.97; RMSE: 1.12; MAE: 0.83; MAPE: 1.70) and testing data ($R^2$: 0.98; RMSE: 1.29; MAE: 2.27; MAPE: 4.08). The model's ability to capture the variability of the system was evaluated using unseen data. The Stk-ANN again had the best performance when predicting on unseen data inside the model boundaries ($R^2$: 0.82; RMSE: 2.57; MAE: 1.75; MAPE: 3.68), suggesting that the Stk-ANN model best captures the system variability within the model boundaries. However, the Stk-ANN was outperformed by the GP and SVR model when extrapolating beyond the model boundaries (COD removal rates > 71%).

Once built, the data-driven models developed using the framework proposed within this study may act as the empirical model within optimisation methodologies (e.g., response surface methodology, evolutionary algorithms) for bioprocess optimisation and intensification. Complications are likely to occur due to bioprocesses' tendency to have multiple, sometimes contradictory, outputs they wish to optimise. For example, the model built within this study predicts the COD removal rate but does not consider the removal rates of other pollutants or the energy generated by the bioprocess during wastewater treatment. Therefore, future work should investigate the implications of developing multiple output models using the proposed framework and how multiple outputs for bioprocess optimisation and intensification.

## Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161/1], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing [grant number EP/P001246/1]; Network Plus: Industrial Systems in a Digital Age.

## CRediT authorship contribution statement

**Oliver J. Fisher**: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Nicholas J. Watson**: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Laura Porcu**: Investigation, Validation, Resources. **Darren Bacon**: Validation, Resources. **Martin Rigley**: Validation, Resources. **Rachel L. Gomes**: Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The author (RLG) would like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K014161], Cloud Manufacturing – Towards Resilient and Scalable High Value Manufacturing. The author (OJF) would like to acknowledge the University of Nottingham, Faculty of Engineering for his PhD scholarship. The author (OJF) would also like to acknowledge the support of Engineering and Physical Sciences Research Council (EPSRC) Network Plus: Industrial Systems in a Digital Age [grant number EP/P001246/1], Connected Everything for funding his placement at Lindhurst Engineering Ltd.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bej.2022.108499.

## References

[1] P. Noll, M. Henkel, History and evolution of modeling in biotechnology: modeling & simulation, application and hardware performance, Comput. Struct. Biotechnol. J. 18 (2020) 3309–3323, https://doi.org/10.1016/J.CSBJ.2020.10.018.

[2] K. Bernaerts, J.F.Van Impe, Data-driven approaches to the modelling of bioprocesses, Trans. Inst. Meas. Control. 26 (2004) 349–372, https://doi.org/10.1191/0142331204tm127oa.

[3] D. Solomatine, L.M. See, R.J. Abrahart, Data-driven modelling: concepts, approaches and experiences, in: R.J. Abrahart, L.M. See, D.P. Solomatine (Eds.), Pract. Hydroinformatics. Water Sci. Technol. Libr, Springer, Berlin, 2008, pp. 17–31, https://doi.org/10.1007/978-3-540-79881-1.

[4] C.W. Coley, W.H. Green, K.F. Jensen, Machine learning in computer-aided synthesis planning, Acc. Chem. Res. 51 (2018) 1281–1289, https://doi.org/10.1021/acs.accounts.8b00087.

[5] P. Bruce, A. Bruce, Practical Statistics for Data Scientists, third ed., O'Reilly Media, Sebastopol, 2018.

[6] S.M. Al-Fattah, H.A. Al-Naim, H.A. Al-Naim, Artificial-intelligence technology predicts relative permeability of giant carbonate reservoirs, SPE Reserv. Eval. Eng. 12 (2009) 96–108, https://doi.org/10.2118/109018-PA.

[7] S.J. Qin, Process data analytics in the era of big data, AIChE J. 60 (2014) 3092–3100, https://doi.org/10.1002/aic.14523.

[8] O.J. Fisher, N.J. Watson, J.E. Escrig, R. Witt, L. Porcu, D. Bacon, M. Rigley, R. L. Gomes, Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems, Comput. Chem. Eng. (2020), 106881, https://doi.org/10.1016/j.compchemeng.2020.106881.

[9] A. Borgogna, A. Salladini, L. Spadacini, A. Pitrelli, M.C. Annesini, G. Iaquaniello, Methanol production from Refuse Derived Fuel: influence of feedstock composition on process yield through gasification analysis, J. Clean. Prod. (2019), https://doi.org/10.1016/J.JCLEPRO.2019.06.185.

[10] G. Garcia-Garcia, J. Stone, S. Rahimifard, Opportunities for waste valorisation in the food industry – a case study with four UK food manufacturers, J. Clean. Prod. 211 (2019) 1339–1356, https://doi.org/10.1016/J.JCLEPRO.2018.11.269.

[11] D. Rodriguez-Granrose, A. Jones, H. Loftus, T. Tandeski, W. Heaton, K.T. Foley, L. Silverman, Design of experiment (DOE) applied to artificial neural network architecture enables rapid bioprocess improvement, Bioprocess Biosyst. Eng. 44 (2021) 1301–1308, https://doi.org/10.1007/S00449-021-02529-3/FIGURES/4.

[12] V. Abt, T. Barz, N. Cruz, C. Herwig, P. Kroll, J. Möller, R. Pörtner, R. Schenkendorf, Model-based tools for optimal experiments in bioprocess engineering, Curr. Opin. Chem. Eng. 22 (2018) 244–252, https://doi.org/10.1016/J.COCHE.2018.11.007.

[13] K. Ažman, J. Kocijan, Application of Gaussian processes for black-box modelling of biosystems, ISA Trans. 46 (2007) 443–457, https://doi.org/10.1016/J.ISATRA.2007.04.001.

[14] E. Hlangwani, W. Doorsamy, J.A. Adebiyi, L.I. Fajimi, O.A. Adebo, A modeling method for the development of a bioprocess to optimally produce umqombothi (a South African traditional beer), Sci. Rep. 11 (2021) 1–15, https://doi.org/10.1038/s41598-021-00097-w.

[15] J. Möller, K.B. Kuchemüller, T. Steinmetz, K.S. Koopmann, R. Pörtner, Model-assisted Design of Experiments as a concept for knowledge-based bioprocess development, Bioprocess Biosyst. Eng. 42 (2019) 867–882, https://doi.org/10.1007/S00449-019-02089-7/FIGURES/7.

[16] J. Panerati, M.A. Schnellmann, C. Patience, G. Beltrame, G.S. Patience, Experimental methods in chemical engineering: artificial neural networks–ANNs, Can. J. Chem. Eng. 97 (2019) 2372–2382, https://doi.org/10.1002/cjce.23507.

[17] S. Kim, MATLAB Deep learning with machine learning, neural networks and artificial intelligence, 1st ed., Apress, 2017. doi:10.1007/978–1-4842–2845-6.

[18] S. Bajracharya, M. Sharma, G. Mohanakrishna, X. Dominguez, D.P.B.T.B Strik, P. M. Sarma, X. Dominguez Benneton, D. Pant, An overview on emerging bioelectrochemical systems (BESs): technology for sustainable electricity, waste remediation, resource recovery, chemical production and beyond, Renew. Energy 98 (2016) 153–170, https://doi.org/10.1016/j.renene.2016.03.002.

[19] A. Donoso-Bravo, J. Mailier, C. Martin, J. Rodríguez, C.A. Aceves-Lara, A. V. Wouwer, Model selection, identification and validation in anaerobic digestion: a review, Water Res. 45 (2011) 5347–5364, https://doi.org/10.1016/j.watres.2011.08.059.

[20] O.J. Fisher, N.J. Watson, L. Porcu, D. Bacon, M. Rigley, R.L. Gomes, Multiple target data-driven models to enable sustainable process manufacturing: an industrial bioprocess case study, J. Clean. Prod. 296 (2021), 126242, https://doi.org/10.1016/j.jclepro.2021.126242.

[21] G. Albuquerque, T. Lowe, M. Magnor, Synthetic generation of high-dimensional datasets, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2317–2324, https://doi.org/10.1109/TVCG.2011.237.

[22] I. Hobæk Haff, K. Aas, A. Frigessi, On the simplified pair-copula construction - simply useful or too simplistic? J. Multivar. Anal. 101 (2010) 1296–1310, https://doi.org/10.1016/j.jmva.2009.12.001.

[23] F.P. Brissette, M. Khalili, R. Leconte, Efficient stochastic generation of multi-site synthetic precipitation data, J. Hydrol. 345 (2007) 121–133, https://doi.org/10.1016/J.JHYDROL.2007.06.035.

[24] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, F. Herrera, A survey on data preprocessing for data stream mining: current status and future directions, Neurocomputing 239 (2017) 39–57, https://doi.org/10.1016/J.NEUCOM.2017.01.078.

[25] W. Yang, K. Wang, W. Zuo, Neighborhood component feature selection for high-dimensional data image enhancement and restoration, J. Comput. 7 (2012) 161–168, https://doi.org/10.4304/jcp.7.1.161-168.

[26] M.H. Abbas, R. Norman, A. Charles, Neural network modelling of high pressure CO2 corrosion in pipeline steels, Process Saf. Environ. Prot. 119 (2018) 36–45, https://doi.org/10.1016/j.psep.2018.07.006.

[27] H. Liu, M. Cocea, Semi-random partitioning of data into training and test sets in granular computing context, Granul. Comput. 2 (2017) 357–386, https://doi.org/10.1007/s41066-017-0049-2.

[28] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[29] X. Zeng, G. Luo, Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection, Heal. Inf. Sci. Syst. 5 (2017) 1–21, https://doi.org/10.1007/s13755-017-0023-z.

[30] D. Gómez, A. Rojas, An empirical overview of the no free lunch theorem and its effect on real-world machine learning classification, Neural Comput. 28 (2015) 216–228, https://doi.org/10.1162/NECO_a_00793.

[31] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205, https://doi.org/10.1016/J.RSER.2017.04.095.

[32] Y. Liu, R. Gunawan, Bioprocess optimization under uncertainty using ensemble modeling, J. Biotechnol. 244 (2017) 34–44, https://doi.org/10.1016/J.JBIOTEC.2017.01.013.

[33] M. Dalmau, N. Atanasova, S. Gabarrón, I. Rodriguez-Roda, J. Comas, Comparison of a deterministic and a data driven model to describe MBR fouling, Chem. Eng. J. (2015), https://doi.org/10.1016/j.cej.2014.09.003.

[34] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

[35] P. Mathews, Design of xxperiments language and concepts, in: Des. Exp. With MINITAB, ASQ Quality Press, Milwaukee, 2004, pp. 93–142.

[36] A.M. Schweidtmann, A.D. Clayton, N. Holmes, E. Bradford, R.A. Bourne, A. A. Lapkin, Machine learning meets continuous flow chemistry: automated optimization towards the Pareto front of multiple objectives, Chem. Eng. J. 352 (2018) 277–282, https://doi.org/10.1016/J.CEJ.2018.07.031.

[37] C.G. Piuleac, M.A. Rodrigo, P. Cañizares, S. Curteanu, C. Sáez, Ten steps modeling of electrolysis processes by using neural networks, Environ. Model. Softw. (2010), https://doi.org/10.1016/j.envsoft.2009.07.012.

[38] A. Hemmati-Sarapardeh, M.-H. Ghazanfari, S. Ayatollahi, M. Masihi, Accurate determination of the $CO_2$-crude oil minimum miscibility pressure of pure and impure $CO_2$ streams: a robust modelling approach, Can. J. Chem. Eng. 94 (2016) 253–261, https://doi.org/10.1002/cjce.22387.

[39] D. Güçlü, Ş. Dursun, Artificial neural network modelling of a large-scale wastewater treatment plant operation, Bioprocess Biosyst. Eng. 33 (2010) 1051–1058, https://doi.org/10.1007/s00449-010-0430-x.

[40] M.N. Kashani, S. Shahhosseini, A methodology for modeling batch reactors using generalized dynamic neural networks, Chem. Eng. J. (2010), https://doi.org/10.1016/j.cej.2010.02.053.

[41] C. Kelleher, T. Wagener, Ten guidelines for effective data visualization in scientific publications, Environ. Model. Softw. 26 (2011) 822–827, https://doi.org/10.1016/j.envsoft.2010.12.006.

[42] Intelligent Plant, (n.d.). ⟨https://www.intelligentplant.com/index.html⟩ (Accessed 2 July 2019).

[43] AnalytiQs - Data Visualisation For Manufacturing - Valuechain, (n.d.). ⟨https://valuechain.com/analytiqs⟩ (Accessed 13 August 2019).

[44] D.B. Carr, R.J. Littlefield, W.L. Nicholson, J.S. Littlefield, Scatterplot matrix techniques for large, New J. Am. Stat. Assoc. 82 (1987) 424–436, https://doi.org/10.1080/01621459.1987.10478445.

[45] K. Wang, A. Salhi, E.S. Fraga, Process design optimisation using embedded hybrid visualisation and data analysis techniques within a genetic algorithm optimisation framework, Chem. Eng. Process. Process. Intensif. 43 (2004) 657–669, https://doi.org/10.1016/J.CEP.2003.01.001.

[46] F. Charte, I. Romero, M.D. Pérez-Godoy, A.J. Rivera, E. Castro, Comparative analysis of data mining and response surface methodology predictive models for enzymatic hydrolysis of pretreated olive tree biomass, Comput. Chem. Eng. (2017), https://doi.org/10.1016/j.compchemeng.2017.02.008.

[47] T. Hastie, R. Tibshirani, J. Friedman, Overview of supervised learning, Elem. Stat. Learn (2009) 9–41.

[48] K. Singh, Dharmendra, Power density analysis by using soft computing techniques for microbial fuel cell, Microbial. Fuel Cell 7 (2019) 1068–1073.

[49] C. Wenfang, K. Lesnik, M. Wade, E. Heidrich, Y.-H. Wang, H. Liu, Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells, Biosens. Bioelectron. 133 (2019) 64–71, https://doi.org/10.1016/j.bios.2019.03.021.

[50] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006 (Accessed 15 August 2019), ⟨http://www.gaussianprocess.org/gpml/⟩.

[51] J. Wang, Q. Wang, J. Zhou, X. Wang, L. Cheng, Operation space design of microbial fuel cells combined anaerobic–anoxic–oxic process based on support vector regression inverse model, Eng. Appl. Artif. Intell. 72 (2018) 340–349, https://doi.org/10.1016/j.engappai.2018.04.005.

[52] A. Garg, V. Vijayaraghavan, S.S. Mahapatra, K. Tai, C.H. Wong, Performance evaluation of microbial fuel cell by artificial intelligence methods, Expert Syst. Appl. 41 (2014) 1389–1399, https://doi.org/10.1016/J.ESWA.2013.08.038.

[53] A. de Ramón-Fernández, M.J. Salar-García, D. Ruiz Fernández, J. Greenman, I. A. Ieropoulos, Evaluation of artificial neural network algorithms for predicting the effect of the urine flow rate on the power performance of microbial fuel cells, Energy 213 (2020), 118806, https://doi.org/10.1016/j.energy.2020.118806.

[54] M. Esfandyari, M.A. Fanaei, R. Gheshlaghi, M.A. Mahdavi, Neural network and neuro-fuzzy modeling to investigate the power density and Columbic efficiency of microbial fuel cell, J. Taiwan Inst. Chem. Eng. 58 (2016) 84–91, https://doi.org/10.1016/J.JTICE.2015.06.005.

[55] K. Larson Lesnik, H. Liu, Predicting microbial fuel cell biofilm communities and bioreactor performance using artificial neural networks, Environ. Sci. Technol. (2017), https://doi.org/10.1021/acs.est.7b01413.

[56] G.E.P. Box, Science and statistics, J. Am. Stat. Assoc. 71 (1976) 791–799. ⟨http://mkweb.bcgsc.ca/pointsofsignificance/img/Boxonmaths.pdf⟩ (accessed August 22, 2019).

[57] M.S. Zaghloul, R.A. Hamza, O.T. Iorhemen, J.H. Tay, Comparison of adaptive neuro-fuzzy inference systems (ANFIS) and support vector regression (SVR) for data-driven modelling of aerobic granular sludge reactors, J. Environ. Chem. Eng. 8 (2020), 103742, https://doi.org/10.1016/j.jece.2020.103742.

[58] K.M. Desai, S.A. Survase, P.S. Saudagar, S.S. Lele, R.S. Singhal, Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: case study of fermentative production of scleroglucan, Biochem. Eng. J. 41 (2008) 266–273, https://doi.org/10.1016/j.bej.2008.05.009.

[59] M. Yolmeh, M.B. Habibi Najafi, F. Salehi, Genetic algorithm-artificial neural network and adaptive neuro-fuzzy inference system modeling of antibacterial activity of annatto dye on Salmonella enteritidis, Microb. Pathog. 67–68 (2014) 36–40, https://doi.org/10.1016/j.micpath.2014.02.003.

[60] D.L.J. Alexander, A. Tropsha, D.A. Winkler, Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, J. Chem. Inf. Model. 55 (2015) 1316–1322, https://doi.org/10.1021/acs.jcim.5b00206.

[61] A.-N. Spiess, N. Neumeyer, An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach, BMC Pharm. 10 (2010) 6, https://doi.org/10.1186/1471-2210-10-6.

[62] M. Welling, Product of experts, Scholarpedia 2 (2007) 3879, https://doi.org/10.4249/SCHOLARPEDIA.3879.

[63] European Union, Directive 2010/75/EU of the European Parliament and of the Council of 24 November 2010 on industrial emissions (integrated pollution prevention and control), OJ L 334, 2010.

[64] X. Zhu, X. Wang, Y.S. Ok, The application of machine learning methods for prediction of metal sorption onto biochars, J. Hazard. Mater. 378 (2019), 120727, https://doi.org/10.1016/j.jhazmat.2019.06.004.