**Ensemble Learning for bioprocess dynamic modelling and prediction**

Max R. Mowbraya, Ehecatl A. Del Rio-Chanonab, Irina Haruna, Jonathan L. Wagnerc, Klaus Hellgardtb, Dongda Zhanga,b,\*

aDepartment of Chemical Engineering and Analytical Science, the University of Manchester, Oxford Road, Manchester, M1 3BU, UK.

bDepartment of Chemical Engineering, Imperial College London, South Kensington, United Kingdom, SW7 2AZ, UK.

cDepartment of Chemical Engineering, Loughborough University, Loughborough, LE11 3TU, UK.

*\*: corresponding author, email:* [*dongda.zhang@manchester.ac.uk*](mailto:dongda.zhang@manchester.ac.uk) *(Dongda Zhang).*

**Abstract**

Machine learning techniques have been successfully used to simulate and optimise bioprocesses. This study explores the feasibility to apply Gradient Boosting, an emerging Ensemble Learning algorithm, which combines weak learners to generate better predictions for bioprocess dynamic modelling and prediction. A thorough procedure was presented for Gradient Boosting based data-driven model construction. Different case studies were employed including fermentation and algal photo-production processes. Given that generating a large size of experimental data for model training is time consuming and challenging to many bioprocesses, this work launched a first investigation on the data efficiency of Gradient Boosting by comparing its predictive capability against the predominantly used artificial neural networks. By carrying out a series of experimental verifications over a broad spectrum of process operating conditions, this study concluded that Gradient Boosting may have several advantages in small experimental datasets and can outperform artificial neural networks for bioprocess predictive modelling, indicating its potential for future bioprocess digitalisation and optimisation.

**Keywords:** Gradient Boosting; artificial neural network; dynamic modelling; data-driven modelling; excreted biofuel.

**1. Introduction**

Industrially focused mathematical modelling technologies have been extensively used for the design, optimisation, control, and scale-up of various bio-production processes. At present, using biological resources such as microalgae and cyanobacteria for producing and processing materials and chemicals has been identified as a global strategic plan to reduce the dependency on petrochemicals and achieve a low carbon economy (Manirafasha, Ndikubwimana, Zeng, Lu, & Jing, 2016; Vavitsas, Fabris, & Vickers, 2018). A number of biosynthetic routes have been developed to produce commercial biorenewables used in the chemical, energy, food, cosmetics, and pharmaceutical industries by utilising organic wastes, CO2 and solar energy as the feedstock (Davies, Work, Beliaev, & Posewitz, 2014; Manirafasha et al., 2018). However, in contrast to industrial chemical processes which are operated under steady-state conditions, bioprocesses are predominantly operated dynamically (*e.g.* fed-batch operation) and periodically (Del Rio-Chanona, Cong, Bradford, Zhang, & Jing, 2018). In addition, bioprocesses are governed by complex metabolic mechanisms characterised by unknown natural patterns and low reproducibility (Vatcheva, de Jong, Bernard, & Mars, 2006a). As a result, it is particularly challenging to establish high-accuracy mathematical models to simulate and predict the behaviours of a bioprocess.

Traditionally, bioprocess modelling is accomplished by constructing a predictive kinetic model modified from several classic kinetic models such as the Monod model, the Droop model, and the Logistic model (Vatcheva, de Jong, Bernard, & Mars, 2006b). These predictive kinetic models have demonstrated excellent efficiency in simulating and optimising fermentation processes. They have been successfully exploited to design and scale up a number of bioprocesses for the industrialisation of different compounds (Jing et al., 2018; Wagner, Lee-Lane, Monaghan, Sharifzadeh, & Hellgardt, 2019; D. Zhang, Chanona, Vassiliadis, & Tamburic, 2015). Furthermore, their applications have also been extended to simulate more complex systems *e.g.* algal and cyanobacterial photo-production processes by incorporating other physical models tackling light transmission and fluid dynamics (Cho & Pott, 2019; Dongda Zhang, Dechatiwongse, & Hellgardt, 2015). However, their predictive ability has been found low in these cases due to the lack of prior knowledge of the underlying systems (Adesanya, Davey, Scott, & Smith, 2014; Del Rio-Chanona et al., 2018). In addition, identifying the suitable mathematical structure of a kinetic model is time consuming and often presents structural problems relevant to parameter estimation and identifiability (Bernard, Dochain, Genovesi, Gouze, & Guay, 2008).

As a result, machine learning based models have been widely studied over the last decade to overcome the above challenges. In particular, feedforward neural networks, one of the most commonly used artificial neural networks (ANNs), have been adopted to simulate chemical and biochemical systems operated under different scales (del Rio-Chanona et al., 2018; Peng et al., 2014), and have been applied to the chemical industry for process control (Baughman & Liu, 1995). ANNs have also exhibited excellent predictive power when optimising complex biosystems, and have yielded the highest productivity on several valuable biomaterials (del Rio-Chanona, Manirafasha, Zhang, Yue, & Jing, 2016; Dineshkumar, Dhanarajan, Dash, & Sen, 2015). From 2018, another machine learning technique, Gaussian processes (GPs) which automatically estimate process uncertainty, have received significant attention for model-plant mismatch estimation, bioprocess monitoring, dynamic optimisation, and product quality control (Bradford, Schweidtmann, Zhang, Jing, & del Rio-Chanona, 2018; Tulsyan, Garvin, & Ündey, 2018). Nonetheless, due to the lack of physical knowledge support, most data-driven models rely on the availability of substantial datasets obtained from a wide range of operating conditions, which could be very time consuming. Such prerequisite can even be infeasible when dealing with photo-production processes where microalgae and cyanobacteria are cultivated to produce biorenewables via photosynthesis, as these strains grow much slower than bacteria or yeast and each process may take up to months. Thus, it is worthwhile to investigate other machine learning methods, which could be potentially more data efficient.

In particular, the current study aims to explore the applicability of an emerging Ensemble Learning technique, Gradient Boosting, for bioprocess modelling and prediction. Ensemble Learning is a class of machine learning methods, which combines a number of weak learners to generate better predictions (a stronger learner) (Yang, Hwa Yang, B. Zhou, & Y. Zomaya, 2010). This work uses decision trees as the weak learners given their high interpolation but poor extrapolation effectiveness (Guido & Mueller, 2016). Gradient Boosting has achieved state-of-the-art performance across a wide variety of tasks such as multi-class classification, data mining, and learning-to-rank (*e.g.* website ranking) (Ke et al., 2017; Natekin & Knoll, 2013). It has been used to test quantitative structure-activity relationships for the pharmaceutical industry in 2016 (Sheridan, Wang, Liaw, Ma, & Gifford, 2016) and estimate catalyst effectiveness by the chemical industry in 2018 (Mistry, Letsios, Krennrich, Lee, & Misener, 2018). However, its efficiency in dynamic bioprocess modelling is yet to be tested. As a result, this work aims to launch the first investigation on its potential in bioprocess predictive modelling and digitalisation, and compare its performance against artificial neural networks, the most widely used data-driven model.

**2. Methodology**

**2.1 Introduction to Gradient Boosting**

Gradient Boosting (GB) is a subclass of Ensemble Learning methods. The technique used in GB to connect individual weak models and improve model accuracy is named *boosting*, in which a new predictor is learned sequentially to correct the mistake made by its previous predictor (Satopää & Veaux, 2011). In the case of a GB model, construction is such that the decision trees inside are trained in a stepwise, sequential fashion to minimise the net error generated from their predecessor, as illustrated in Fig. 1(a). For process regression and prediction, boosting is found to outperform *bagging* – another ensembling technique that enhances model accuracy by collecting and averaging results from individual independent decision trees (*e.g.* Random Forest which is predominantly used in classification as shown in Fig. 1(b)) (Guido & Mueller, 2016).

The development of GB is inspired by the steepest descent method in optimisation (Nocedal & Wright, 2006). In mathematical optimisation, steepest descent method (shown as Eq. (1a)) searches a local minimum of an objective function by taking steps along the opposite direction to the gradient of the function at the current point, as gradient represents the direction of greatest increase of the function (Ganjisaffar, Caruana, & Lopes, 2011). GB incorporates this concept for its construction. Assume that GB consists of decision trees. The training data has input , where each element is a vector containing different features, and output , where each element is a single value. The GB is a Multiple-Input and Single-Output (MISO) model. Using the concept of steepest descent, the output of the *k+1th* decision tree is calculated by Eq. (1b), where the term approximates the gradient of a loss function at which minimises the residue between GB’s prediction and . Both and are learned during the training process until the model is constructed. A more detailed training algorithm has been described by Satopää and Veaux (Satopää & Veaux, 2011).

where and are the point found at the *k+1th* and *kth* step, respectively, is the step length at the *kth* step, and is the gradient of an objective function at . and are the output of the *k+1th* and *kth* decision tree, respectively, is the weight associated with the *kth* tree, is learning rate, and .

Decision trees in GB have an “if-else” structure to partition data into different groupings (Mistry et al., 2018). Each of the terminal nodes (leaves) of a tree represents a cell of the partition, and contains a simple model that only applies to the specific leaf. Traditionally, each leaf in a decision tree is assigned to a single value, meaning that all the training samples entering the same leaf are given the same prediction. These trees are called piecewise constant trees, resulting in a piecewise linear GB model (Shi, Li, & Li, 2018). Such GBs are easy to construct and have shown excellent prediction accuracy to various tasks. For more complex systems, recent advances have embedded linear and nonlinear equations into the leaves to improve model efficiency (Shi et al., 2018).

Compared to other widely used machine learning methods *e.g.* ANN and GP, the advantages of GB include that its training algorithm is invariant to scaling of the data and that it is efficient to learn from datasets with a small size (Guido & Mueller, 2016). However, its accuracy is sensitive to several hyperparameters including number of trees , and maximum depth of each tree (or maximum number of leaves per tree) and learning rate (Ganjisaffar et al., 2011). Learning rate in Eq. (1c) is a parameter used to modify the weight (step length) for the boosting update. It controls how strongly each tree tries to correct the mistake made by its predecessor (Natekin & Knoll, 2013). Decreasing learning rate can reduce the risk of model overfitting. Whilst smaller learning rates require a higher number of boosting iterations, resulting in a larger model with an increased number of trees (Ganjisaffar et al., 2011). Although the addition of trees can lead to a better training result, it once again increases the risk of overfitting. To avoid this trade-off, GB often uses shallow trees (*e.g.* depth less than five) for regression (Guido & Mueller, 2016). The optimal combination of these hyperparameters was thoroughly explored in this study utilising a variance-bias framework as detailed in the subsequent sections.

**2.2 GB model structure identification**

The identification of a proper model structure is fundamental to the quality of any estimator, irrespective of whether it is mechanistic or data-driven. Mechanistic model structure identification has a strong basis in process knowledge, whereas determination of the topology and parameters defining a data-driven model poses a complex nonlinear, combinatorial optimisation problem where gradients are typically unavailable. Hyperparameter optimisation of data-driven models has been well reported in the literature, with the emergence and growth of the automated machine learning community primarily driven by wider-industries, particularly Google and Amazon (He, Zhao, & Chu, 2019). In demonstrating the utility of GB trees for the prediction of dynamic bioprocess systems, this work implements a direct search or simplex method, known as Nelder-Mead. The work utilises this approach in view of its ease of implementation and the computational demands associated with manual and exhaustive grid search procedures. The following details further advantages and implementation of the algorithm.

The Nelder-Mead method acts to sample an *n* dimensional search space via the maintenance and transformation of a non-degenerate *n* + 1 dimensional simplex. A number of conditional rules guide this transformation with activation based on the relative values of each vertex of the simplex as evaluated by the objective function. At each iteration, appropriate activation of a rule instigates either a reflection, contraction, expansion or shrinkage, enabling the simplex to traverse the search space. The optimisation procedure terminates upon the satisfaction of a criterion, typically a relative bound on transformation of the simplex to inhibit the algorithm from implementing too small a step size (Bartholomew–Biggs, 2008; Kolda, Lewis, & Torczon, 2003). A more detailed description of the implementation of this algorithm is presented in Algorithm 1.

Despite wide use within many engineering fields, this method is unable to guarantee convergence upon global optima and is reported to terminate at minimax points within high-dimensional search spaces. Detailed review of the Nelder-Mead algorithm, in the context of other direct search methods, is provided by Lewis(Lewis, Torczon, & Trosset, 2000)*.* Nevertheless, the method is appropriate for this work given the low computational demand, relatively low-dimensionality of the search space and ease of implementation. Model selection was made via the akaike information criterion (AIC),which facilitates bias-variance analysis and identification of parsimonious model structures via balance of a parameter penalty and model error (Konishi, 2008; Von Stosch, Peres, de Azevedo, & Oliveira, 2010). The AIC is defined by Eq. (2a) to Eq. (2c):

where is the number of trees, is the maximum depth of each tree, is the number of model parameters and is the number of data points.

|  |
| --- |
| **Algorithm 1** Nelder-Mead Algorithm for Model Selection |
| **Input:** Initialise:where – in this work (see Eq (1c) and (2a) for notation); the parameters of transformation, to satisfy constraints as outlined in (Rao, 2009) – this work implements values as detailed in the standard form of the algorithm (Lagarias, Reeds, Wright, & Wright, 1998); the stopping criteria tolerance.  **Output:** Best point and  **for** j = 1, …, **Termination** **do**   1. **Order** the set such that , where is the iteration of the algorithm and check **Termination** 2. Calculate , where 3. Perform **Reflection:**; if, remove , accept (*i.e.* consolidate the new vertex) and return to Step 1. 4. Perform **Expansion:** if then **.** If accept; else accept. Remove and return to Step 1. 5. Perform **Contraction**: if , then . If , accept , replace and return to Step 1. 6. Perform **Shrink**: for all , where , . Return to Step 1. 7. Perform **Termination:** if , where |

**2.3 Stochastic gradient boosting trees**

As GB is sensitive to its hyperparameters, to further prevent overfitting, another recommended strategy is to introduce robustness into the training procedure (Friedman, 2002). Otherwise known as stochastic GB, the approach selects a random subset of the available training data to train each sequential tree. The technique not only provides regularisation, but also increases the speed and accuracy of model training. The technique introduces greater variance in individual tree predictions, but provides a variance reduction in overall model prediction leading to greater accuracy (Friedman, 2002; Hastie, 2009)**.** Typically the size of the subset is in the region of 50% of the available data, but may be tuned depending on the task and number of data points (Hastie, 2009). In this work, smaller subsets produced improved performance for tasks where the predicted variable displayed little change between data points (*i.e.* biomass), with larger subsets preferred when the converse was observed (*i.e.* in the case of nitrate prediction). In other words, when the change of predicted variable fluctuates considerably, models with less variance between the predictions of constituent trees were favoured.

**2.4 Construction of Artificial Neural Network (ANN) model**

ANNs, in specific feedforward neural networks, are the primary machine learning technology used in chemical process modelling. Its training procedure has been well documented in several publications (Baughman & Liu, 1995; del Rio-Chanona et al., 2017), hence not repeated in this work. Although other types of neural networks such as recurrent neural networks may provide better data fitting result for time-series events from a theoretical perspective and have been also applied to bioprocess modelling during early 2000s (Barrera-Cortés, Baruch, Valdez-Castro, & Vázquez-Cervantes, 2001; Chen, Nguang, Chen, & Li, 2004), a number of more recent works have shown that in practice a well-trained feedforward neural network is accurate enough to simulate most dynamic bioprocesses (del Rio-Chanona et al., 2016, 2018; Dineshkumar et al., 2015; do Carmo Nicoletti & Jain, 2009; Peng et al., 2014). In addition, recurrent neural networks may require more datasets for training given their complex model structure – this is particularly disadvantageous when the investigated system only has limited useful data as commonly experienced in the biochemical industry (typically unknown as “small data”). As a result, feedforward neural network is chosen as the first benchmark in this work to compare against GB. Key hyperparameters for ANN construction such as the number of hidden layers, the number of neurons, and learning rate were identified through a hyperparameter selection framework comprising a *k*-fold method and a variance-bias analysis detailed in our recent work (Del Rio-Chanona et al., 2019), hence not repeated here.

**2.5 Data augmentation**

Despite the simplicity of ANNs and GBs, they may still require the availability of more datasets than a mechanistic model for training (parameter estimation), as a data-driven model can easily contain hundreds of parameters whilst a theory-driven bioprocess model usually has less than 20 parameters. As a result, a data augmentation strategy proposed in our previous research was adopted here to generate substantial artificial datasets based on a small set of original experimental data whenever needed (del Rio-Chanona et al., 2016). This strategy efficiently creates a large size of artificial data by embedding an adequate random noise (often 3%-5%) into the original data to represent the highly regulated yet still stochastic nature of bioprocesses, and has been found to be effective in a number of applications for bioprocess modelling, optimisation, and monitoring (Bradford et al., 2018; del Rio-Chanona et al., 2018; Tulsyan et al., 2018). Effects of this strategy on GB construction are discussed in the results section.

**2.5 Training of data-driven models**

As GB is a Multiple-Input and Single-Output (MISO) model, a number of independent GBs are constructed to constitute a GB framework, each of which estimates the change of concentration of a specific state variable. It is worth noting that this approach was initially developed for Gaussian processes (GP) regression (Bradford et al., 2018), another MISO machine learning technique, which has been previously adopted for biosystem simulation. In contrast, ANN is a Multiple-Input and Multiple-Output (MIMO) model, so only one ANN is constructed to simulate all state variables. Model training was executed using the *scikit-learn* software library (Pedregosa et al., 2012) in Python 3.6 and was terminated upon identification of the optimal combination of hyperparameters. Simulation of mechanistic models utilised Mathematica 11.

**3. Introduction to case studies**

In this work, two case studies are used to test the applicability of GB for bioprocess predictive modelling, and their details are explained here.

**3.1 Fed-batch fermentation process**

In the first case study, a simple fermentation kinetic model was used as a computational experiment for initial investigation. The kinetic model is shown by Eq. (3a) and Eq. (3b). The system describes biomass growth and nitrate consumption within a fed-bath operation where nitrate is continuously replenished into the culture:

where and are biomass and nitrate concentrations, respectively. This is a simple case study by design, which allows enough simplicity for 2D visualisation and parameter analysis. A more complex case study is presented later in this work.

Details regarding the values of kinetic parameters are listed in Supporting Information (Table S1). The use of a computational case study enables the generation of many datasets displaying varied dynamics depending on the control strategy implemented, in other words, the temporal trajectory of (nitrate inflow rate, mg L-1 h-1). The motivation of this case study was to provide an initial comparison on the predictive performance of GB and ANN for bioprocess modelling, as well as providing guidance for a better design of data-driven models used in Case study 2.

A total of 14 independent datasets were generated. This was achieved by the variation of three design variables constituting the initial concentration (IC) of states and two factors affecting the trajectory of (*i.e.* the time step to change a control action, and the underlying function). Table 1 specifies the control strategies and initial conditions executed with design of experiments utilising a 2×2×3 full factorial approach. From the datasets available, 9 were used to train the models. The other three datasets (Test sets 1-3) were used to test the data-driven models’ prediction performance when simulating a new experiment operated similar to those used for training data generation. In addition, two further datasets (Test sets 4-5) were generated for the purpose of exploring data-driven models’ predictive capability when the control actions deviate more significantly from previous experiments. Details of the training and test datasets are summarised in Table 2. Each dataset is comprised of 192 data points, equivalent to sampling the system every 2 hours.Fig. 2 provides insight into the computational experiments.

**3.2** **Algal photo-production for excreted biofuel production**

Case study 2 focuses on a more complex bio-production process in which real experimental datasets are directly used for model construction. However, given that each algal photo-production experiment takes over 2 weeks to implement, only a few datasets are available. It is worth emphasising that the lack of experimental data is a common challenge when simulating microalgal photo-production processes. For example, in this study, it takes 6 months to collect 6 training datasets and 5 test datasets due to the much slower algal cells doubling time (~ 26 hours) compared to bacteria (~30 min) (Gibson, Wilson, Feil, & Eyre-Walker, 2018; Liu, Huang, & Chen, 2011). Thus, due to the significant time cost for experimentation, identifying a data-driven model that is able to effectively learn from “small data” is of particular importance. Microalgal bisabolene production is used in this case study. Bisabolene is a novel biofuel which can be directly excreted to the culture after synthesis (Harun et al., 2018). It is recognised as an ideal jet fuel alternative (precursor to bisabolane, a D2-diesel fuel) and has huge global market demand (Wichmann, Baier, Wentnagel, Lauersen, & Kruse, 2018). Developing a digital tool to enable the predictive modelling and scale-up of this system is of great importance to the energy and biochemical industries.

Production of bisabolene is greatly affected by light intensity and temperature. As a result, these two factors are selected as the design variables. In total, 11 batch experiments were implemented across different light and temperature conditions as listed in Table 3. Six datasets (Training datasets 1-6) covering a wide range of light intensity and a narrow range of temperature were used to train the data-driven models, whereas the remaining 5 datasets (Tests 1-5) were used to check the model’s predictive capabilities. In particular, operating conditions of Tests 1-2 are similar to the training experiments, whereas Tests 3-5 deviate from the training experiments substantially (different temperatures). The purpose of designing different tests is to explore the reliability of data-driven models when predicting an unknown process. Most datasets include concentrations of biomass, nitrate (substrate), and bisabolene every 12 hours until the end of the experiment (144 hours), except for Tests 3-5 in which nitrate concentrations were not measured.

**4. Results and discussion**

**4.1 Results of data-driven modelling for fermentation process**

As previously detailed, GB is invariant to data scaling and efficient to learn from small-sized datasets. In practice, data scaling (normalisation) is easy to implement and has been well established. Therefore, it is more important to test if GB is more data efficient for bioprocess predictive modelling. Therefore, two GB models were designed in Case study 1: ***GB-I:*** without data augmentation; and ***GB-II:*** with data augmentation.

Optimal combinations of key hyperparameters of the stochastic GB and ANN models are identified and summarised in Table 4. It is observed that data augmentation increases the complexity of the ANN structure by embedding 50% of extra neurons into each hidden layer in this work. One must note that the most powerful effect of data augmentation is the mitigation of overfitting when only scarce datasets are available. However, the effect of data augmentation on GB model structure is not so evident, although it indeed affects the total number of trees in the model. This is notable in the case of biomass prediction (the number of trees nearly doubled), but it is less pronounced in nitrate prediction. Such conclusion is likely attributed to the model’s ‘if-else’ structure and categorical partitioning of data.

As stated before, subsampling is a technique introduced to facilitate rapid training and improve the accuracy of GB models. The greater the subsampling fraction, the larger the subset of training data required to train each tree. Interestingly, there is clear difference between the GB models favoured subsampling fraction. The performance of GB biomass models improves with a lower subsampling fraction, whilst contrarily, GB nitrate models show disposition for a higher subsampling fraction. This is invariant of data augmentation and normalisation. It indicates that a greater portion of the data is required for nitrate model training, which is likely caused by the dramatic change of nitrate concentration due to frequent changes of nitrate inflow control actions. Both the ANN and the GB models can well represent the training datasets as shown in Fig. 3(a) and (b), indicating the successful construction of these data-driven models.

A further calculation (Supporting Information Table S2) shows that data augmentation does not significantly improve GB models’ prediction. In fact, there is marginal increase in the performance of GBs without data augmentation. This is probably because the original data is already sufficiently rich in information (1728 training data points from different operations) to train the GB models, hence adding artificial data only increases the complexity of model configuration as previously detailed. It suggests that compared to an ANN, GB’s, for this case study, accuracy is not necessarily to be improved by data augmentation in view of the increased complexity of model structure, therefore it is more data efficient. Moreover, as shown in Fig. 3 (c)-(f), the predictive performance of the GB and ANN models in Case study 1 are comparable across the majority of test experiments. In most cases, the two models’ prediction results are visually inseparable. Based on calculation (listed in Supporting Information Table S3), the GB models are found to slightly outperforms the ANN for all the tests, showing its potential in simulating nonlinear time-series biochemical processes.

**4.2 Results of data-driven modelling for algal photo-production process**

To further confirm the competitiveness of GB in applications subject to small datasets, Case study 2 is carried out. Compared to Case study 1, due to the practical limitation, much fewer training experimental data points are available in this case (48 points from the original experiments relative to 1728 points in Case study 1). Thus, artificial data becomes essential to assist construction of the data-driven models, and mitigate overfitting to the scarce datapoints available. In this work, 300 artificial datasets (2400 artificial data points) were created using the method introduced in Section 2.5 to train the data-driven models. It is, however, important to clarify that artificial datasets do not contribute extra biological knowledge to the data-driven models; instead, they are mainly used to assist the data-driven models to learn more efficiently from (*i.e.* extract useful information) the original experimental datasets.

Fig. 4 shows the training and prediction performance of the GB and ANN models in Case study 2. Table 5 summarises the prediction errors of the two types of models. From the table, it is seen that in most cases the GB and the ANN models have comparable accuracy when predicting different experiments. This is also confirmed in Fig. 4 (b), (d), and (f) where both models’ simulation results are difficult to distinguish. However, large prediction errors beyond 30% are observed when the two data-driven models are used to simulate Tests 3-5. Temperatures in these experiments (22-26 °C) were set far away from that used in the training experiments (28-32 °C). As cell growth and product synthesis are sensitive to the narrow change of temperature whilst the training data does not include this information, it is observed that data-driven models are not suited to predict an unknown process in which the operating conditions differ significantly from previous experiments.

In fact, when the operations of new processes do not change greatly from the previous training experiments (Tests 1-2), both GB and ANN can show accurate predictions with an error mostly less than 5%, with GB yielding slightly better predictions than ANN. However, when the unknown process is operated significantly different from the previous experiments (Tests 3-4), GB yields the least reliable predictions, introducing errors substantially larger than ANN for predicting biomass growth, as would be expected with data-driven models. Nonetheless, these errors reduce dramatically (*e.g.* decreased by 10 times) when the new process’ operating conditions move close to GB’s training data (*e.g.* temperature in Test 5 is 26 °C), making GB a more accurate predictive model than ANN. This phenomenon further demonstrates that great caution should be given when using machine learning models to predict process behaviour, as their prediction may not be reliable beyond their training operational region and that they should not be used to simulate (extrapolate) an unknown process in which the operating conditions significantly deviate from the training experiments.

**5. Conclusions**

Overall, this work shows that Gradient Boosting can be considered as a competent alternative to artificial neural networks for bioprocess modelling. It is an efficient machine learning method that can well learn from small experimental datasets (with the support of artificial data argumentation if necessary), which is particularly attractive for bioprocess dynamic simulation and predictive modelling. When estimating an unknown process in which the operating conditions do not deviate substantially from the training experiments, GB may outperform ANN and give more accurate results. However, when the new process is conducted far away from the conditions of the training experiments, it is not recommended to rely on purely a machine learning method. As a result, future work should consider how to integrate process biological and physical knowledge into data-driven modelling techniques so that their predictive power can be significantly extended.

**6. Acknowledgement**

This project has received funding from the EPSRC project (EP/P016650/1, P65332). This project has also received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 640720. Irina acknowledges funding from the Ministry of Higher Education, Malaysia and University Putra Malaysia.

**7. References**

Adesanya, V. O., Davey, M. P., Scott, S. A., & Smith, A. G. (2014). Kinetic modelling of growth and storage molecule production in microalgae under mixotrophic and autotrophic conditions. *Bioresource Technology*, *157*, 293–304. https://doi.org/10.1016/j.biortech.2014.01.032

Barrera-Cortés, J., Baruch, I., Valdez-Castro, L., & Vázquez-Cervantes, V. (2001). A Recurrent Neural Network Model of the B.t. Fed-Batch Fermentation Process 1. *IFAC Proceedings Volumes*, *34*(5), 385–390. https://doi.org/10.1016/S1474-6670(17)34250-7

Bartholomew–Biggs, M. (2008). Direct Search Methods. In *Nonlinear Optimization with Engineering Applications* (pp. 1–10). https://doi.org/10.1007/978-0-387-78723-7\_5

Baughman, D. R., & Liu, Y. A. (1995). *Neural Networks in Bioprocessing and Chemical Engineering*. https://doi.org/10.1016/C2009-0-21189-5

Bernard, O., Dochain, D., Genovesi, A., Gouze, J.-L., & Guay, M. (2008). *Bioprocess Control* (D. Dochain, Ed.). https://doi.org/10.1002/9780470611128

Bradford, E., Schweidtmann, A. M., Zhang, D., Jing, K., & del Rio-Chanona, E. A. (2018). Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. *Computers & Chemical Engineering*, *118*, 143–158. https://doi.org/10.1016/j.compchemeng.2018.07.015

Chen, L., Nguang, S. K., Chen, X. D., & Li, X. M. (2004). Modelling and optimization of fed-batch fermentation processes using dynamic neural networks and genetic algorithms. *Biochemical Engineering Journal*, *22*(1), 51–61. https://doi.org/10.1016/j.bej.2004.07.012

Cho, B. A., & Pott, R. W. M. (2019). The development of a thermosiphon photobioreactor and analysis using Computational Fluid Dynamics (CFD). *Chemical Engineering Journal*, *363*, 141–154. https://doi.org/10.1016/j.cej.2019.01.104

Davies, F. K., Work, V. H., Beliaev, A. S., & Posewitz, M. C. (2014). Engineering Limonene and Bisabolene Production in Wild Type and a Glycogen-Deficient Mutant of Synechococcus sp. PCC 7002. *Frontiers in Bioengineering and Biotechnology*, *2*. https://doi.org/10.3389/fbioe.2014.00021

Del Rio-Chanona, E. A., Ahmed, N. R., Wagner, J., Lu, Y., Zhang, D., & Jing, K. (2019). Comparison of physics‐based and data‐driven modelling techniques for dynamic optimisation of fed‐batch bioprocesses. *Biotechnology and Bioengineering*, *116*(11), 2971–2982. https://doi.org/10.1002/bit.27131

Del Rio-Chanona, E. A., Cong, X., Bradford, E., Zhang, D., & Jing, K. (2018). Review of advanced physical and data‐driven models for dynamic bioprocess simulation: Case study of algae–bacteria consortium wastewater treatment. *Biotechnology and Bioengineering*, bit.26881. https://doi.org/10.1002/bit.26881

del Rio-Chanona, E. A., Fiorelli, F., Zhang, D., Ahmed, N. rashid, Jing, K., & Shah, N. (2017). An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process. *Biotechnology and Bioengineering*, *114*(11), 2518–2527. https://doi.org/10.1002/bit.26373

del Rio-Chanona, E. A., Manirafasha, E., Zhang, D., Yue, Q., & Jing, K. (2016). Dynamic modeling and optimization of cyanobacterial C-phycocyanin production process by artificial neural network. *Algal Research*, *13*, 7–15. https://doi.org/10.1016/j.algal.2015.11.004

del Rio-Chanona, E. A., Wagner, J. L., Ali, H., Fiorelli, F., Zhang, D., & Hellgardt, K. (2018). Deep learning‐Based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design. *AIChE Journal*, aic.16473. https://doi.org/10.1002/aic.16473

Dineshkumar, R., Dhanarajan, G., Dash, S. K., & Sen, R. (2015). An advanced hybrid medium optimization strategy for the enhanced productivity of lutein in Chlorella minutissima. *Algal Research*, *7*, 24–32. https://doi.org/10.1016/j.algal.2014.11.010

do Carmo Nicoletti, M., & Jain, L. C. (Eds.). (2009). *Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control*. https://doi.org/10.1007/978-3-642-01888-6

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Ganjisaffar, Y., Caruana, R., & Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information - SIGIR ’11*, 85. https://doi.org/10.1145/2009916.2009932

Gibson, B., Wilson, D. J., Feil, E., & Eyre-Walker, A. (2018). The distribution of bacterial doubling times in the wild. *Proceedings. Biological Sciences*, *285*(1880), 20180789–20180789.

Guido, S., & Mueller, A. C. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media.

Harun, I., Del Rio-Chanona, E. A., Wagner, J. L., Lauersen, K. J., Zhang, D., & Hellgardt, K. (2018). Photocatalytic Production of Bisabolene from Green Microalgae Mutant: Process Analysis and Kinetic Modeling. *Industrial & Engineering Chemistry Research*, *57*(31), 10336–10344. https://doi.org/10.1021/acs.iecr.8b02509

Hastie, Trevor. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition.). New York, NY: Springer New York.

He, X., Zhao, K., & Chu, X. (2019). AutoML: A Survey of the State-of-the-Art. *ArXiv.Org*. Retrieved from http://search.proquest.com/docview/2268454117/

Jing, K., Tang, Y., Yao, C., del Rio-Chanona, E. A., Ling, X., & Zhang, D. (2018). Overproduction of L-tryptophan via simultaneous feed of glucose and anthranilic acid from recombinant Escherichia coli W3110: Kinetic modeling and process scale-up. *Biotechnology and Bioengineering*, *115*(2), 371–381. https://doi.org/10.1002/bit.26398

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems*, 9. NIPS Proceedings.

Kolda, T. G., Lewis, R. M., & Torczon, V. (2003). Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods. *SIAM Review*, *45*(3), 385,482.

Konishi, Sadanori. (2008). *Information Criteria and Statistical Modeling*. New York, NY: Springer New York.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions. *SIAM Journal on Optimization*, *9*(1), 112,147.

Lewis, R. M., Torczon, V., & Trosset, M. W. (2000). Direct search methods: Then and now. *Journal of Computational and Applied Mathematics*, *124*(1), 191–207. https://doi.org/10.1016/S0377-0427(00)00423-4

Liu, J., Huang, J., & Chen, F. (2011). Microalgae as Feedstocks for Biodiesel Production. In M. Stoytcheva & G. Montero (Eds.), *Biodiesel*. https://doi.org/10.5772/25600

Manirafasha, E., Murwanashyaka, T., Ndikubwimana, T., Rashid Ahmed, N., Liu, J., Lu, Y., … Jing, K. (2018). Enhancement of cell growth and phycocyanin production in Arthrospira (Spirulina) platensis by metabolic stress and nitrate fed-batch. *Bioresource Technology*, *255*, 293–301. https://doi.org/10.1016/j.biortech.2017.12.068

Manirafasha, E., Ndikubwimana, T., Zeng, X., Lu, Y., & Jing, K. (2016). Phycobiliprotein: Potential microalgae derived pharmaceutical and biological reagent. *Biochemical Engineering Journal*, *109*, 282–296. https://doi.org/10.1016/j.bej.2016.01.025

Mistry, M., Letsios, D., Krennrich, G., Lee, R. M., & Misener, R. (2018). *Mixed-Integer Convex Nonlinear Optimization with Gradient-Boosted Trees Embedded*.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Nocedal, Jorge., & Wright, S. (2006). *Numerical Optimization* (Second Edition.). New York, NY: Springer New York.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2012). Scikit-learn: Machine Learning in Python. *CoRR*, *abs/1201.0490*. Retrieved from http://arxiv.org/abs/1201.0490

Peng, W., Zhong, J., Yang, J., Ren, Y., Xu, T., Xiao, S., … Tan, H. (2014). The artificial neural network approach based on uniform design to optimize the fed-batch fermentation condition: Application to the production of iturin A. *Microbial Cell Factories*, *13*(1), 54. https://doi.org/10.1186/1475-2859-13-54

Rao, S. S. (2009). *Engineering optimization: Theory and practice* (Fourth edition.). Hoboken, NJ: John Wiley & Sons.

Satopää, V. A., & Veaux, R. D. De. (2011). *A Robust Boosting Algorithm for Chemical Modeling*. Philadelphia.

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, *56*(12), 2353–2360. https://doi.org/10.1021/acs.jcim.6b00591

Shi, Y., Li, J., & Li, Z. (2018). *Gradient Boosting With Piece-Wise Linear Regression Trees*.

Tulsyan, A., Garvin, C., & Ündey, C. (2018). Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. *Biotechnology and Bioengineering*, *115*(8), 1915–1924. https://doi.org/10.1002/bit.26605

Vatcheva, I., de Jong, H., Bernard, O., & Mars, N. J. I. (2006a). Experiment selection for the discrimination of semi-quantitative models of dynamical systems. *Artificial Intelligence*, *170*(4–5), 472–506. https://doi.org/10.1016/j.artint.2005.11.001

Vatcheva, I., de Jong, H., Bernard, O., & Mars, N. J. I. (2006b). Experiment selection for the discrimination of semi-quantitative models of dynamical systems. *Artificial Intelligence*, *170*(4–5), 472–506. https://doi.org/10.1016/j.artint.2005.11.001

Vavitsas, K., Fabris, M., & Vickers, C. (2018). Terpenoid Metabolic Engineering in Photosynthetic Microorganisms. *Genes*, *9*(11), 520. https://doi.org/10.3390/genes9110520

Von Stosch, M., Peres, J., de Azevedo, S., & Oliveira, R. (2010). Modelling biochemical networks with intrinsic time delays: A hybrid semi-parametric approach. *BMC Systems Biology*, *4*(1), 131.

Wagner, J. L., Lee-Lane, D., Monaghan, M., Sharifzadeh, M., & Hellgardt, K. (2019). Recovery of excreted n-butanol from genetically engineered cyanobacteria cultures: Process modelling to quantify energy and economic costs of different separation technologies. *Algal Research*, *37*, 92–102. https://doi.org/10.1016/j.algal.2018.11.008

Wichmann, J., Baier, T., Wentnagel, E., Lauersen, K. J., & Kruse, O. (2018). Tailored carbon partitioning for phototrophic production of ( E )-α-bisabolene from the green microalga Chlamydomonas reinhardtii. *Metabolic Engineering*, *45*, 211–222. https://doi.org/10.1016/j.ymben.2017.12.010

Yang, P., Hwa Yang, Y., B. Zhou, B., & Y. Zomaya, A. (2010). A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*, *5*(4), 296–308. https://doi.org/10.2174/157489310794072508

Zhang, D., Chanona, E. A. D.-R., Vassiliadis, V. S., & Tamburic, B. (2015). Analysis of green algal growth via dynamic model simulation and process optimization. *Biotechnology and Bioengineering*, *112*(10). https://doi.org/10.1002/bit.25610

Zhang, Dongda, Dechatiwongse, P., & Hellgardt, K. (2015). Modelling light transmission, cyanobacterial growth kinetics and fluid dynamics in a laboratory scale multiphase photo-bioreactor for biological hydrogen production. *Algal Research*, *8*(0), 99–107. https://doi.org/10.1016/j.algal.2015.01.006