An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process

Ehecatl Antonio del Rio-Chanona[1,2], Fabio Fiorelli[1], Dongda Zhang[2,*], Nur rashid Ahmed[3], Keju Jing[3], Nilay Shah[2]

1: Department of Chemical Engineering and Biotechnology, University of Cambridge, Pembroke Street, Cambridge CB2 3RA, UK.

2: Centre for Process Systems Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.

3. Department of Chemical and Biochemical Engineering, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China.

*: corresponding authors, email: dongda.zhang11@imperial.ac.uk, tel: 44 (0)7543785283.

Running title: Dynamic Modelling of Algal Lutein Production

Abstract

Lutein is a high-value bioproduct synthesised by microalga *Desmodesmus* sp.. It has great potential for the food, cosmetics, and pharmaceutical industries. However, in order to enhance its productivity and to fulfil its ever-increasing global market demand, it is vital to construct accurate models capable of simulating the entire behaviour of the complicated dynamics of the underlying biosystem. To this aim, in this study two highly robust artificial neural networks are designed for the first time. Contrary to conventional artificial neural networks, these networks model the rate of change of the dynamic system, which makes them highly relevant in practice. Different strategies are incorporated into the current research to guarantee the accuracy of the constructed models, which include determining the optimal network structure through a hyper-parameter selection framework, generating significant amounts of artificial data sets by embedding random noise of appropriate size, and rescaling model inputs through standardisation. Based on experimental verification, the high accuracy and great predictive power of the current models for long-term dynamic bioprocess simulation in both real-time and offline frameworks are thoroughly demonstrated. This research, therefore, paves the way to significantly facilitate the future investigation of lutein bioproduction process control and optimisation. In addition, the model construction strategy developed in this research has great potential to be directly applied to other bioprocesses.

Keywords: artificial neural network; dynamic simulation; lutein production; real-time framework; fed-batch operation; bioprocess modelling.

1. Introduction

The synthesis of valuable bioproducts from microalgae through photosynthetic related metabolic pathways is a promising sector (Mata et al. 2010) that can deliver a wide variety of commodities. One of the most well-known applications is the production of eco-friendly biofuels such as biodiesel and biohydrogen (Doshi et al. 2016; Zhang, Dechatiwongse, del Rio-Chanona, et al. 2015), which are being developed to replace traditional transportation fuels. Another use is that as food supplements (Río et al. 2005; Xie et al. 2014), a capability in which they have been long employed and for which considerable growth of market demand is forecasted. Most importantly, there is a thriving international research interest, development and deployment of microalgae based sustainable and environmentally friendly technologies, by the health sector. This is with the aim of producing specialist high-value products such as lutein (Xie et al. 2014), C-phycocyanin (del Rio-Chanona, Zhang, et al. 2015), and astaxanthin (Zhang et al. 2016), for which traditional synthesis routes and refinery methods from existing non-renewable sources are expensive, energy intensive, and of low efficiency (Yen et al. 2011; Capelli et al. 2013).

One high-value bioproduct that has found particular attention is lutein, a carotenoid of great interest to the health, pharmaceutical, cosmetics, and food industries (Yaakob et al. 2014; Fernández-Sevilla et al. 2010). Lutein has been widely used for the treatment of ophthalmic conditions and cancer, and applied as a natural colorant in cellular pigmentation and in the food industry(Ho et al. 2015; Xie et al. 2013). Because of its wide applications, its market demand in the US was estimated to increase significantly from $150 million in 2000 to $309 million in 2018 (Marz 2011). Nonetheless, further growth in production is hampered by the fact that its current industrial feedstock is marigold, a plant which has an extremely low intracellular content of

lutein (0.02~0.1% wt (fresh flowers)) with low growth rate (Lin et al. 2015), explaining the current high process separation costs and low productivity.

Therefore, microalgae with their rapid growth rate and ability to utilise plenty of low cost and abundant resources including solar energy, atmospheric CO_2 and wastewater, can provide significant improvements over existing industrial practice. Research into microalgal lutein production has already made considerable progress. It was found that nitrogen intake into the system is vital for the synthesis and accumulation of lutein (Xie et al. 2014; Ho et al. 2015). It was also discovered that low illumination is a decisive factor to ensure the suitable conditions for lutein production (Ho et al. 2012; Xie et al. 2013). Furthermore, the influence of different photobioreactor types and operating modes on microalgae growth and lutein synthesis were studied (Del Campo 2000; del Rio-Chanona, Zhang, et al. 2016). Different microalgae species such as *Scenedesmus obliquus*, *Chlorella sorokiniana*, *Chlorella zofingiensis* and *Desmodesmus* sp. were found to produce lutein with an intercellular lutein content that is between 6 and 15 times higher than marigold (Shi et al. 2002; Sánchez et al. 2008; Del Campo 2000).

Despite these achievements, one of the challenges that most severely prevents the industrialisation of the microalgae based lutein production process is to efficiently conduct a dynamic optimisation of the process. Precise control over a long-term bioprocess is indispensable to guarantee its safety and efficiency, as bioprocesses are very sensitive to the process operating conditions such as pH, temperature and nutrient supply. Hence, executing process optimisation within a control scheme can remarkably improve the process profitability (del Rio-Chanona, Zhang, et al. 2016). In order to resolve this challenge, it is essential to initially

construct a robust mathematical model, which is able to accurately simulate and predict the dynamic performance of the long-term photo-production system, for both microalgae growth and lutein production throughout the entire process (Zhang, Dechatiwongse, Del-Rio-Chanona, et al. 2015).

However, so far little attention has been paid on this aspect. At present, two methodologies have been predominantly developed for bioprocess simulation, namely kinetic modelling and artificial neural networks (ANNs). Specific to microalgal lutein production, a kinetic model including effects of both nutrient concentration and light intensity on biomass growth and lutein synthesis has not been proposed in the literature. In addition, although a kinetic model can display good accuracy and predictability, it is noticed that constructing such a complex model is always a difficult mathematical task (*e.g.* parameter estimation of highly nonlinear ordinary differential equation systems) even if the bioproduct synthesis mechanisms have been identified. Hence, its application in bioprocess control and real-time optimization has not yet been well conducted.

On the contrary, ANNs have been widely used in traditional chemical engineering processes for process control (Mjalli 2005; Fissore et al. 2004), and their applications have been recently extended to biochemical systems (Witek-Krowiak et al. 2014; Rosales-Colunga et al. 2010). Furthermore, they have been successfully employed for the purpose of reproducing, controlling, and optimising the dynamical behaviour of microalgae based bioprocesses (García-Camacho et al. 2016; del Rio-Chanona, Manirafasha, et al. 2016). The key advantage of ANN over kinetic modelling is that the investigated system can be treated as a black-box, without the necessity to develop any empirical or analytical correlation. This significantly reduces the difficulty in model

structure design and model parameter estimation. The challenge in constructing ANNs, however, is the requirement of large and varied raw data sets to express good predictions in supervised learning, which is particularly time consuming for long-term bioprocesses (Witek-Krowiak et al. 2014). As a result, there is not much development in the use of ANN for longer lasting biological systems (del Rio-Chanona, Manirafasha, et al. 2016).

Therefore, to facilitate the industrialisation of lutein production and investigate the capability of ANNs in long-term bioprocess simulation, the work presented here aims to construct robust ANNs capable of accurately modelling and predicting the dynamic behaviour of microalgae biomass growth and lutein synthesis. Different strategies will be adopted to resolve the challenge arising from the limited amount of experimental data sets. In particular, *Desmodesmus* sp. is selected in the current study due to its highest intracellular lutein content (up to 5.0 mg/g) and great thermo-tolerant properties (highest growth rate at 35℃ and can survive up to 46℃) compared to other algae species (Xie et al. 2014; Xie et al. 2013).

The work is divided in the following sections. In this Section 1, a background introduction has been provided to both the object of investigation and the main tools that are used later in this study. In Section 2, the specific details of the implementation of ANNs in the current work are laid out, including a schematic of the structure of the ANNs and use of the “elbow rule” in deciding the key parameters. In Section 3, the results are presented and discussed, showing how with a small amount of experimental data points the designed ANNs can give robust predictions of the dynamic behaviours of the current investigated bioprocesses. This demonstrates their suitability for use in both real-time and offline optimisation frameworks. Finally, in the

conclusion section, the original findings are summarised and an overview of future avenues of research is provided.

2. Theory and methods

2.1 Experiment setup

Six fed-batch experimental processes have been carried out in our work. In these experiments, microalga *Desmodesmus* sp. F51 was used for lutein production, and the operating temperature was controlled at 35℃ for biomass growth and lutein synthesis. This is because 35℃ has been reported to be the optimal temperature for *Desmodesmus* sp. F51 growth (Xie et al. 2013). A 1 L tubular photobioreactor (15.5 cm in length and 9.5 cm in diameter) was used in these experiments with an external light source mounted on both sides. Initial biomass concentrations in these experiments were kept constant, and incident light intensities were set from 150 μmol m^{-2} s^{-1} to 600 μmol m^{-2} s^{-1}. Nitrate influent was supplied to the reactor from the 60^{th} hour until the end of the process due to the consumption of initial culture nitrate, and its inflow rate was fixed at 3 mL hr^{-1}. Influent nitrate concentration was chosen as 0.1 M or 0.5 M in different experiments, and all the processes lasted for six days.

Biomass concentration, nitrate concentration, and lutein production in all the processes were measured once per 12 hours, thus in total for each experimental data set there are 12 data points. Amongst these data sets, four of them were used in the current study for ANN training (model construction), and the remaining two (Test 1 and Test 2, under different light intensity and inflow rate conditions) were used for the predictability verification of the ANNs. All of the experiments were in duplicate, and the detailed presentation of the experimental measurement

techniques can be found in the previous research (Xie et al. 2013). The detailed operating conditions of these experiments are listed in Table I.

2.2 Selection of number of hidden layers

In this study, two different ANN structures were explored and their performance tested. A first ANN with one hidden layer, and a second ANN with two hidden layers. This enabled us to determine which structure is the most suitable one for future process optimization and control, given that few theoretical bases would highlight one over the other. Both ANN structures were fully linked and implemented in Python 2.7 using *pybrain* as a library (Schaul et al. 2010). The function implemented in the hidden layer nodes was chosen to be a sigmoid function, with a linear function in the output layer. This choice comes from recommendations in the literature that suggest such a combination is robust and flexible (Nielsen 2014).

2.3 Inputs and outputs of ANNs

In all cases the ANNs received 5 inputs which are the key operating factors and state variables in the current experiments:

- Incident light intensity;
- Biomass concentration;
- Nitrate concentration;
- Lutein production;
- Nitrate inflow concentration.

Of these inputs, incident light intensity and nitrate inflow concentration are experimentally controllable and known at all times in the experimental settings. The other three inputs are

179 measured once per 12 hours throughout the entire experimental operating time course. All inputs
180 fed to the ANN were rescaled, as the variables in the system under analysis are of very different
181 magnitude and ANNs require similarly scaled inputs. This can benefit the network training in
182 terms of speed and numerical performance, therefore producing a smooth reduction in error
183 during training and promoting a more accurate gradient evaluation.
184
185 Furthermore, the outputs of the ANNs are the change of rate of the state variables including
186 biomass concentration, nitrate concentration, and lutein production. These changes in the rate
187 (either accumulation or consumption) are added into the current ANN inputs to predict the
188 process states at the next time, which is schematically presented in Fig. 1. This strategy was
189 adopted from our recent research in which using the change of states by giving their past
190 information is found to give the network higher accuracy compared to directly predicting future
191 states based on past ones (del Rio-Chanona, Manirafasha, et al. 2016).
192
193 **2.4 Selection of number of neurons in hidden layers**
194 Through literature correlations between the number of inputs and outputs desired, an initial
195 number of nodes in each hidden layer was estimated to be around 20 (Lawrence et al. 1996;
196 Elisseeff & Paugam-Moisy 1996). The exact number of neurons in the current ANNs, however,
197 was tested through a hyper-parameter selection procedure, together with the number of training
198 epochs (times that the ANN is trained). Four experimental data sets, each with 12 points, were
199 employed in this hyper-parameter selection step.
200
201 This hyper-parameter selection was conducted in *k*-fold fashion, where a selection of 4 of the 5

data sets is used for training and the remaining data set is used to obtain an estimate of the maximum error along the trajectory and across all involved variables. Then another subset of 4 is selected and the resulting network is evaluated against the remaining data set. This procedure is repeated until the specific network parameter configuration has been tested against all the sets, and it is then possible to obtain an average of maximum error for each parameter configuration. These errors are compared, and the network configuration with the lowest error is chosen. The choices for the number of nodes in the hidden layers were 5, 10, 15, 20, and 30, and those for the number of training epochs were 15, 30, 50, 100, 150, 200, and 300. The following figures, *i.e.* Fig. 2 and Fig. 3, show the 3D landscapes representing the intersection of these two choices and the error of cross-validation of the constructed ANNs.

To better select the parameters for the ANNs, a framework known as "elbow rule" is used. The objective of this framework is to select the optimal number of layers and training epochs that would give the best trade-off between accuracy of the network, training time and potential of model over-fitting. While often increasing the number of epochs or layers brings further improvement, there are diminishing returns and the improvement is not worth the increasing training time. In addition, noises from an excessive number of internal parameters can create problems of over-fitting, thus deteriorating the ANN's predictive capabilities.

From Fig. 2 and Fig. 3, it can be seen that for the number of epochs, the error decreases rapidly at the beginning and flattens long before the maximum of 600 training epochs. As a result, 400 epochs were chosen in the current study for both ANNs. In terms of the number of nodes, in the single hidden layer case (Fig. 2) adding more than 20 nodes does not result in an improvement in

performance, while in the ANN with two hidden layers (Fig. 3) a significant increase in the number of nodes beyond 10 seems to even contribute to a decrease in performance. Therefore, 20 nodes were picked for the first case, and 15 nodes per hidden layer were chosen in the second case.

2.5 Training of artificial neural networks

For ANN training, 4 sets of experimental data points, each containing 12 data points, were used. To enhance the model accuracy, 50 replications of artificial data sets were produced based on the original data sets with random noise added of 3 % of the variable size, and a further 50 replications with a 5 % noise. These proportions were selected based on the realistic assessment of the accuracy of the current original experimental measurements. The strategy of embedding adequate random noise into original data sets to generate significant amount of artificial data sets has been found to improve ANNs modelling and predictive power when simulating other biological systems, even when relatively little experimental data is available (del Rio-Chanona, Manirafasha, et al. 2016).

3. Results and discussion

3.1 Training results of the artificial neural networks

Once trained and cross-validated, both ANNs are constructed. The current proposed ANN construction strategy was implemented on a personal computer with low specifications in a realistic time (3 hours for the ANN training). This is more time-efficient than constructing a kinetic model for future process design, since it could be a time-consuming task, in particular at the early research stage, to fully identify the biochemical kinetic mechanisms. It is useful to

mention that modern PCs (*e.g.* Core i7) should have more than acceptable performance (30 minutes at most of ANN training). This is further enhanced by the fact that more advanced ANN libraries allow computation to take place in Graphics Processing Units (GPU) in place of traditional CPU usage. The peculiar architecture of GPUs allows for 10-100 times faster training for equivalent amounts of loss, when CPU computation is used for the same cases.

Fig. 4 and Fig. 5 show the comparison between the training data and the simulation results. In order to assess the effectiveness of the current ANN training procedure, in this case only initial operating conditions are provided, and the ANNs simulate the bioprocess behaviors throughout the 132 hours of operations. For illustrative purpose, the maximum absolute percentage error (MAPE) of the one hidden layer ANN is presented in Fig. 4.

With the above results, it can be clearly seen that the ANNs are capable of accurately modelling the process dynamics to which they have already been exposed, even if only the initial operating conditions are given. Furthermore, if these ANNs are used in process optimization and control, they should be able to predict the performance of the investigated system under the operating conditions which they have never encountered. It is for this reason that two additional experimental data sets obtained from different experiments, namely Test 1 and Test 2, were used to test the predictive power of the designed ANNs.

3.2 Predictability of the ANNs on a real-time framework

During an ongoing experiment, the ANN should be capable of predicting the dynamic behavior of the bioprocess several time steps ahead. For this, a real-time framework can be put into place,

where in future work the system can be either controlled or optimized. After every set number of hours (12 hours in the current work) new experimental measurements would become available and the exact state of the system at that time would be known. Therefore, the ANN would only need to be able to accurately predict the performance of the current lutein production process for 12 hours in advance, as new system data would be available after this time interval.

Given that in the current experimental setting only 12 hours would be needed, both ANNs are used to predict the process behaviors of the two additional experimental tests (Test 1 and Test 2) after 12 hours once a measurement is given, and such prediction is repeated throughout the entire experiment operating time. Fig. 6 and Fig. 7 show the prediction results of the two ANNs for both experimental tests. The MAPE of both models are below 10%, expect for the one hidden layer ANN when predicting nitrate concentration at Experiment Test 2. The results shown in the two figures consist in strong proof that both models can be effectively used for the real-time control and optimization of the investigated bioprocess given their high predictive power.

3.3 Predictability of the ANNs on offline framework

Moreover, to thoroughly explore the feasibility of applying ANNs into an offline optimal control framework where the entire process behavior of an unknown experiment is predicted before its implementation (del Rio-Chanona, Dechatiwongse, et al. 2015), the current ANNs are used to simulate the processes Test 1 and Test 2. It follows the procedure that a single initial experimental point is supplied to the network, then the ANN computes the next state. Nonetheless, at the subsequent time step, instead of using the next experimental measurement as input to the ANN, the last computed simulated point is used. This means that, with the exception

of the first point at *t=0* (initial condition), the ANN is not supplied with any other experimental point. Thus, throughout the entire process, the ANN simulation errors are accumulated and the model to system mismatches are magnified. A competent working model should keep the growth of simulation errors contained within the time horizon investigated.

As an example, Fig. 8 compares the prediction results of the ANN with two hidden layers to the real experimental results. The MAPE of the two hidden layers ANN in both cases are mostly around 7% to 12%. From the figure, it can be appreciated that the prediction of the ANN matches the experimental results of both test sets well. This, as mentioned earlier, shows that the current constructed ANN can predict not only 12 hours in advance, but up to 132 hours with high accuracy. This result strongly indicates the great competence of the current ANN for long-term bioprocess modelling and offline optimization.

The small tendency for a slight consistent error seen in Test 2 might indicate a bias in the training sets, mostly in the concentration of nitrogen, where the increase of concentration influenced by the nitrogen injection is not well represented. This can be attributed to a relative lack of variety in the training sets, as the nitrogen inflow concentration has only two different quantities used as input. This seems to be further compounded by the fact that most differences commence around $t=60^{th}$ hr, when nitrogen flow is switched on. Adding a relevant feature in a model usually decreases the modelling error. However, if the feature describes a variable or quantity in the system that does not vary much, this might actually cause more error than the addition of this feature would correct (Hagan et al. 2014). This could be solved by having more data sets with different nitrogen inflow concentrations as the main point of change. This analysis

317 can therefore support further experimental design.

318

319 **3.4 Comparison of the artificial neural networks**

320 In the current study, the two ANNs are created, one with one hidden layer and the other with two
321 hidden layers, both of which are tested against Test 1 and Test 2. Some comparison results which
322 highlight the difference in performance between the two ANNs are presented in Fig. 9. From the
323 figure, it is observed that the performance of the ANN with two hidden layers is much better than
324 the one with one hidden layer, when simulating the trajectory of the entire process (offline
325 framework). Thus, it is concluded that although both ANNs can provide accurate
326 implementations for a real-time framework, if the aim is to execute offline optimization, only the
327 ANN with two hidden layers should be selected due to its higher accuracy and predictive
328 capacities.

329

330 However, it is important to emphasize that the conclusion of a two hidden layers ANN being
331 superior to a one hidden layer ANN cannot be considered as a general rule for bioprocess
332 modelling and optimization. This is because there exists a trade-off between model accuracy and
333 risk of model over-fitting. In other words, although it is possible to enhance the model fitting
334 results by increasing the number of ANN layers (hence increasing the amount of model
335 parameters), the addition of extra layers can result in an over-fitting to the constructed ANN and
336 severely aggravate the ANN predictive capabilities. Moreover, attention should be also paid on
337 the counter-balance between increasing ANN training times by adding more layers and
338 decreasing returns in improvements of predictive power. Thus, when employing ANNs to
339 simulate an unknown biosystem, it is necessary to adopt the current proposed hyper-parameter

selection framework to determine the optimal ANN structure.

3.5 Strategy of training data rescaling

Furthermore, the current study concluded that when rescaling experimental data points for ANN training, it is vital to choose a suitable rescaling method to guarantee the quality of the network. For example, in the current study, it is found that using the standardization method to center training data on the mean seems to have worked best in scaling the points and obtaining a network with better predictive power. This can indicate that the ANN, when trained, becomes more flexible when data points that maximize the variance are present. A min-max normalization method would have not captured this characteristic as well (Hastie et al. 2009). In addition, this type of scaling displays a problem with the presence of larger outliers, as in such data sets it would force many of the scaled points to lie very close to each other in an attempt to include the outliers, which will inevitably decrease the accuracy of the ANN.

Another advantage of choosing a standardization method is connected to the fact that, in the current study, a sigmoid function was chosen for the hidden layers. Sigmoid functions have the tendency to saturate, meaning that they make little distinction between inputs that are on the extremes of the range and produce the same outputs. These functions have worse learning if the inputs are large for either sign, as the gradients inside the ANN are flattened close to values of 0, which can lead to serious problems in the ANN learning process (Nielsen 2014). A standardization method, however, solves this issue by concentrating inputs in a limited range around a mean.

Conclusions

In the current study, two ANNs were constructed to simulate a long-term dynamic biosystem for microalgae biomass growth and lutein production. To guarantee the high accuracy of the current models, different strategies were implemented during the model construction. These include using a standardization method to rescale inputs into the ANN, adding adequate random noise to efficiently generate sufficient amounts of artificial data sets which are infeasible to obtain in practice, and identifying the optimal ANN structure and parameter values through a hyper-parameter selection framework.

By comparing the ANN simulation results with the training data sets, it is found that the current ANNs can accurately represent the dynamic behavior of the current investigated biosystem. By comparing the ANN prediction results against the two sets of test experimental results, it is concluded that both of the current designed ANNs can be effectively applied into real-time process optimization and control frameworks, which strongly indicates their high potential for future process design and optimization. Furthermore, the current research demonstrated that the ANN with two hidden layers is capable of predicting accurately the entire dynamic trajectory of an unknown process before its implementation, further suggesting its adequateness for its use even in the offline optimization framework and extended process prediction durations.

Moreover, due to the necessity of process modelling and optimization for the design of industrial scale sustainable biochemicals production processes, laboratory scale experimental data provides an important test set and can be used as a starting point for pilot plant tests. Inaccuracies of ANNs can be rapidly and easily corrected by further phases of learning. Thus, it is notable that

the procedure and strategies presented in the current study for ANN construction can be directly transferred to other bioprocesses and make significant contributions to their further industrialization. In addition, effective optimization algorithms (such as stochastic and evolutionary algorithms) can be further developed and embedded into the real-time framework to facilitate the optimization of ANNs for future bioprocess design and optimal control.

Acknowledgments

Author E. A. del Rio-Chanona would like to acknowledge CONACyT scholarship No. 522530 This work was also supported by the National Natural Science Foundation of China (No. 31071488), the National High Technology Research and Development Program 863, China (No.2014AA021701).

References

Del Campo, J., 2000. Carotenoid content of chlorophycean microalgae: factors determining lutein accumulation in Muriellopsis sp. (Chlorophyta). *Journal of Biotechnology*, 76(1), pp.51–59.

Capelli, B., Bagchi, D. & Cysewski, G.R., 2013. Synthetic astaxanthin is significantly inferior to algal-based astaxanthin as an antioxidant and may not be suitable as a human nutraceutical supplement. *Nutrafoods*, 12(4), pp.145–152.

Doshi, A. et al., 2016. Economic and policy issues in the production of algae-based biofuels: A review. *Renewable and Sustainable Energy Reviews*, 64, pp.329–337.

Elisseeff, A. & Paugam-Moisy, H., 1996. Size of multilayer networks for exact learning: analytic approach. In *NIPS'96 Proceedings of the 9th International Conference on Neural*

409 *Information Processing Systems*. MIT Press Cambridge, pp. 162–168.
410 Fernández-Sevilla, J.M., Acién Fernández, F.G. & Molina Grima, E., 2010. Biotechnological
411 production of lutein and its applications. *Applied Microbiology and Biotechnology*, 86(1),
412 pp.27–40.
413 Fissore, D., Barresi, A.A. & Manca, D., 2004. Modelling of methanol synthesis in a network of
414 forced unsteady-state ring reactors by artificial neural networks for control purposes.
415 *Chemical Engineering Science*, 59(19), pp.4033–4041.
416 García-Camacho, F. et al., 2016. Artificial neural network modeling for predicting the growth of
417 the microalga Karlodinium veneficum. *Algal Research*, 14, pp.58–64.
418 Hagan, M.T. et al., 2014. *Neural Network Design* 2nd ed., Martin Hagan.
419 Hastie, T., Tibshirani, R. & Friedman, J., 2009. *The Elements of Statistical Learning*, New York,
420 NY: Springer New York. Available at: http://link.springer.com/10.1007/978-0-387-84858-7.
421 Ho, S.-H. et al., 2015. Effects of nitrogen source availability and bioreactor operating strategies
422 on lutein production with Scenedesmus obliquus FSP-3. *Bioresource Technology*, 184,
423 pp.131–138.
424 Ho, S.-H., Lu, W.-B. & Chang, J.-S., 2012. Photobioreactor strategies for improving the CO2
425 fixation efficiency of indigenous Scenedesmus obliquus CNW-N: Statistical optimization of
426 CO2 feeding, illumination, and operation mode. *Bioresource Technology*, 105, pp.106–113.
427 Lawrence, S., Giles, C.L. & Tsoi, A.C., 1996. *What size neural network gives optimal*
428 *generalization? Convergence properties of backpropagation*,
429 Lin, J.-H., Lee, D.-J. & Chang, J.-S., 2015. Lutein production from biomass: Marigold flowers
430 versus microalgae. *Bioresource Technology*, 184, pp.421–428.
431 Marz, U., 2011. *The Global Market for Carotenoids.*

Mata, T.M., Martins, A.A. & Caetano, N.S., 2010. Microalgae for biodiesel production and other applications: A review. *Renewable and Sustainable Energy Reviews*, 14(1), pp.217–232.

Mjalli, F.S., 2005. Neural network model-based predictive control of liquid–liquid extraction contactors. *Chemical Engineering Science*, 60(1), pp.239–253.

Nielsen, M., 2014. Neural Networks and Deep Learning..

del Rio-Chanona, E.A., Manirafasha, E., et al., 2016. Dynamic modeling and optimization of cyanobacterial C-phycocyanin production process by artificial neural network. *Algal Research*, 13, pp.7–15.

del Rio-Chanona, E.A., Zhang, D., et al., 2015. Dynamic Simulation and Optimization for Arthrospira platensis Growth and C-Phycocyanin Production. *Industrial & Engineering Chemistry Research*, 54(43), pp.10606–10614.

del Rio-Chanona, E.A., Dechatiwongse, P., et al., 2015. Optimal Operation Strategy for Biohydrogen Production. *Industrial & Engineering Chemistry Research*, 54(24), pp.6334–6343.

del Rio-Chanona, E.A., Zhang, D. & Vassiliadis, V.S., 2016. Model-based real-time optimisation of a fed-batch cyanobacterial hydrogen production process using economic model predictive control strategy. *Chemical Engineering Science*, 142, pp.289–298.

Río, E. Del et al., 2005. Efficient one-step production of astaxanthin by the microalgaHaematococcus pluvialis in continuous culture. *Biotechnology and Bioengineering*, 91(7), pp.808–815.

Rosales-Colunga, L.M., García, R.G. & De León Rodríguez, A., 2010. Estimation of hydrogen production in genetically modified E. coli fermentations using an artificial neural network. *International Journal of Hydrogen Energy*, 35(24), pp.13186–13192.

Sánchez, J.F. et al., 2008. Biomass and lutein productivity of Scenedesmus almeriensis: influence of irradiance, dilution rate and temperature. *Applied Microbiology and Biotechnology*, 79(5), pp.719–729.

Schaul, T. et al., 2010. PyBrain. *Journal of Machine Learning Research*, 11, pp.743–746.

Shi, X.-M., Jiang, Y. & Chen, F., 2002. High-Yield Production of Lutein by the Green Microalga Chlorella protothecoides in Heterotrophic Fed-Batch Culture. *Biotechnology Progress*, 18(4), pp.723–727.

Witek-Krowiak, A. et al., 2014. Application of response surface methodology and artificial neural network methods in modelling and optimization of biosorption process. *Bioresource Technology*, 160, pp.150–160.

Xie, Y. et al., 2013. Phototrophic cultivation of a thermo-tolerant Desmodesmus sp. for lutein production: Effects of nitrate concentration, light intensity and fed-batch operation. *Bioresource Technology*, 144, pp.435–444.

Xie, Y.-P. et al., 2014. Simultaneous enhancement of CO2 fixation and lutein production with thermo-tolerant Desmodesmus sp. F51 using a repeated fed-batch cultivation strategy. *Biochemical Engineering Journal*, 86(7), pp.33–40.

Yaakob, Z. et al., 2014. An overview: biomolecules from microalgae for animal feed and aquaculture. *Journal of Biological Research-Thessaloniki*, 21(1), p.6.

Yen, H.-W., Sun, C.-H. & Ma, T.-W., 2011. The Comparison of Lutein Production by Scenesdesmus sp. in the Autotrophic and the Mixotrophic Cultivation. *Applied Biochemistry and Biotechnology*, 164(3), pp.353–361.

Zhang, D., Dechatiwongse, P., Del-Rio-Chanona, E.A., et al., 2015. Analysis of the cyanobacterial hydrogen photoproduction process via model identification and process

simulation. *Chemical Engineering Science*, 128, pp.130–146.

Zhang, D. et al., 2016. Dynamic modelling of Haematococcus pluvialis photoinduction for astaxanthin production in both attached and suspended photobioreactors. *Algal Research*, 13, pp.69–78.

Zhang, D., Dechatiwongse, P., del Rio-Chanona, E.A., et al., 2015. Dynamic modelling of high biomass density cultivation and biohydrogen production in different scales of flat plate photobioreactors. *Biotechnology and Bioengineering*, 112(12), pp.2429–2438.

Table I: Operation conditions of the five experiments. Exp 1~Exp 4 are used for ANN training, and Test 1~Test 2 are used for ANN predictability verification.

Operation conditions	Exp 1	Exp2	Exp3	Exp4	Test 1	Test 2
Initial Biomass g L^{-1}	0.07	0.07	0.07	0.07	0.07	0.07
Initial nitrate con. mM	8.8	30	8.8	8.8	30	8.8
Inflow rate mL h^{-1}	3.0	3.0	3.0	3.0	3.0	3.0
Influent nitrate con. M	0.5	0.5	0.1	0.1	0.5	0.1
Light intensity μmol m^{-2} s^{-1}	300	600	150	600	480	300

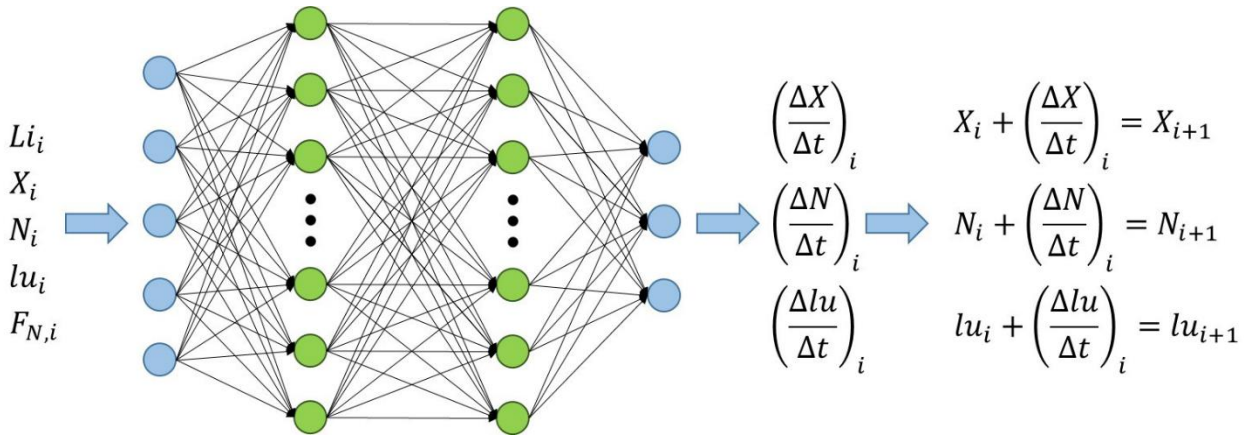


Figure 1: Schematic of the current two hidden layer ANN structure. Time interval is chosen as 12 hours based on the real experimental implementation.

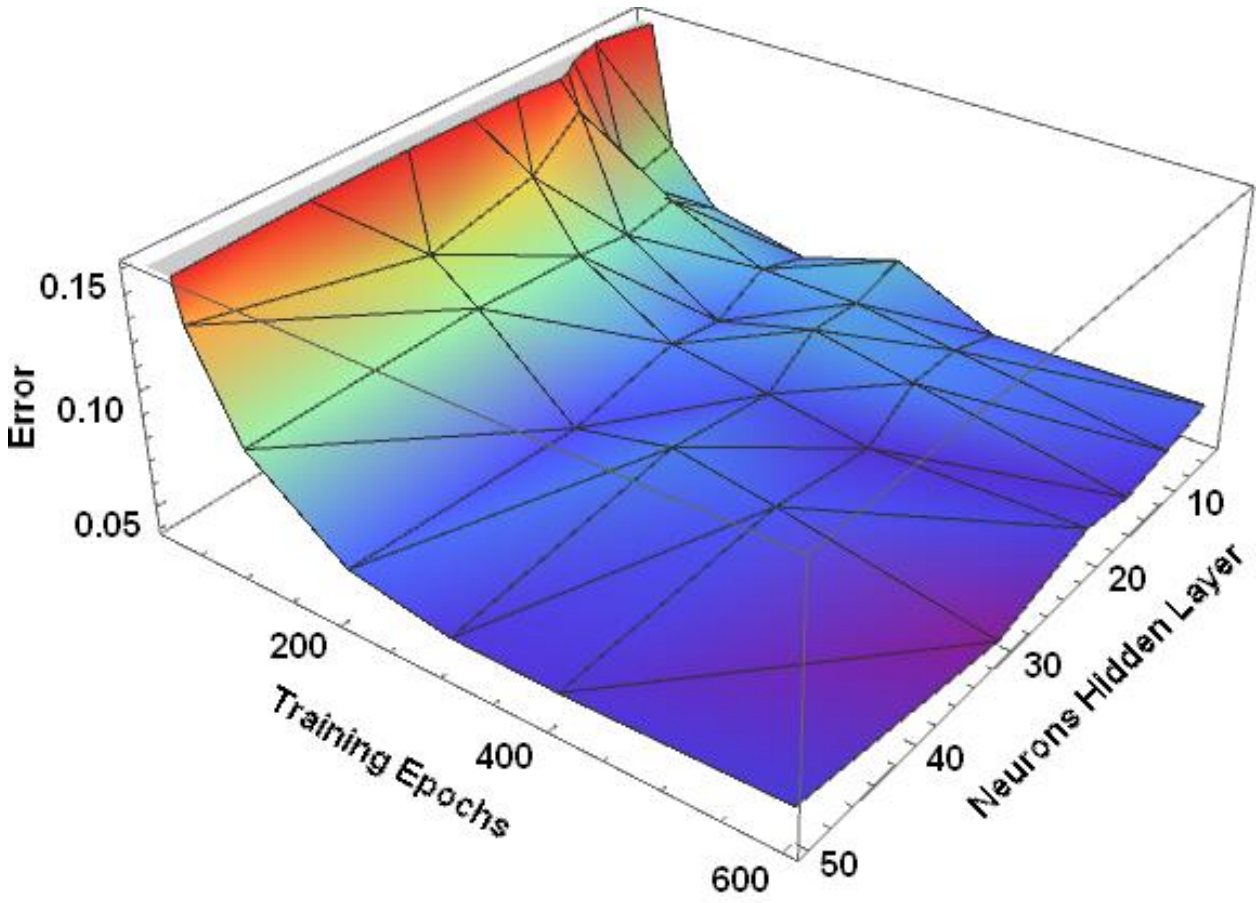


Figure 2: One hidden Layer elbow rule.

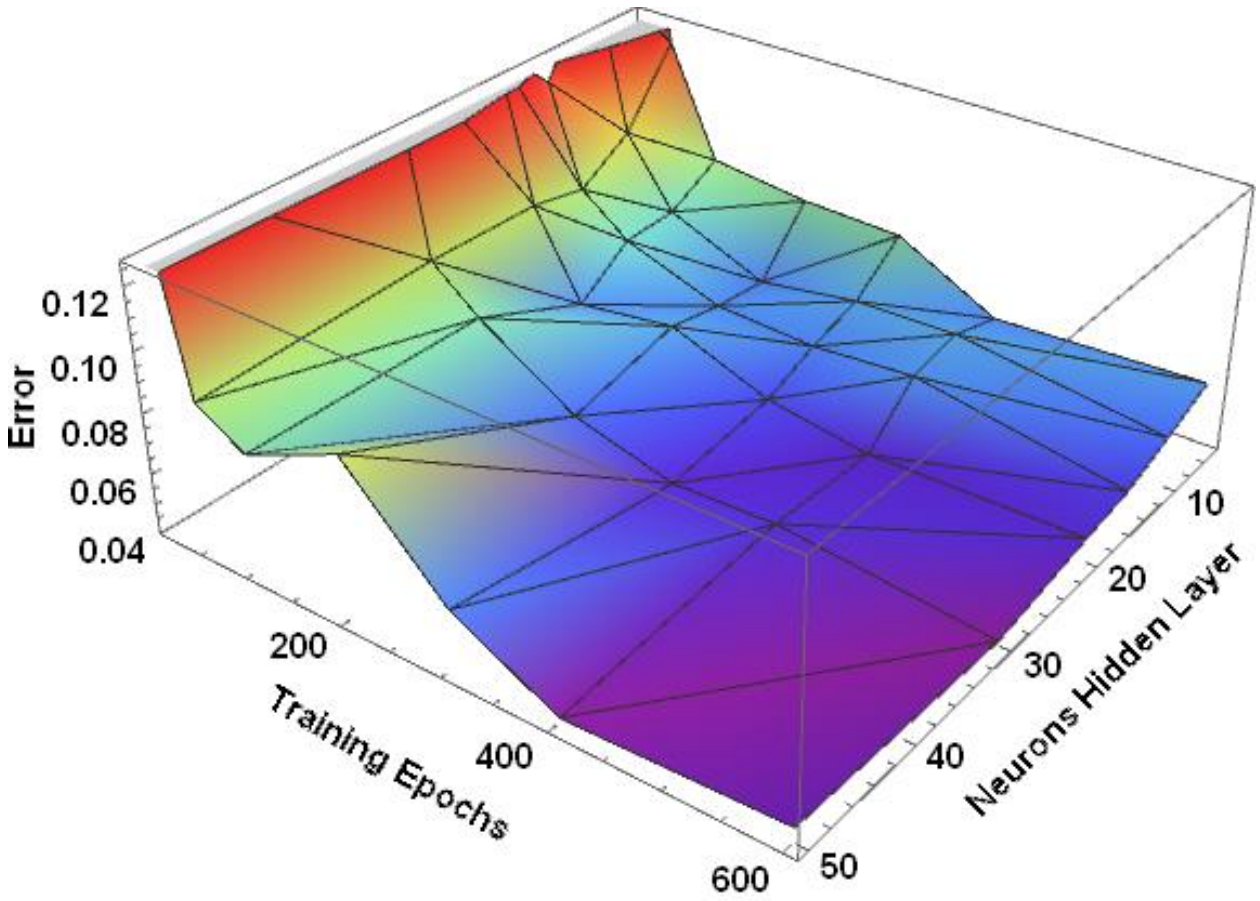


Figure 3: Two hidden Layers elbow rule.

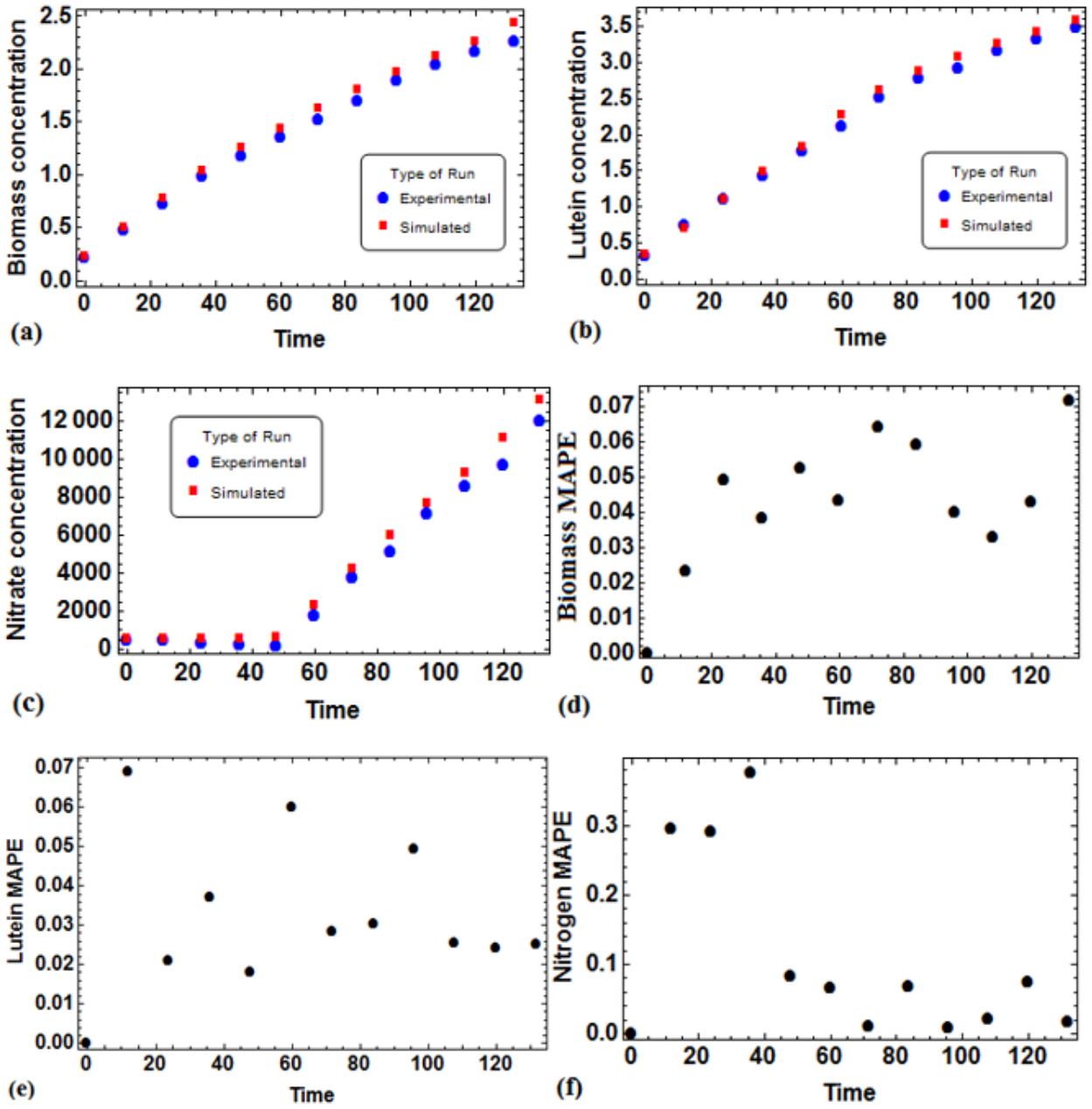


Figure 4: One hidden layer ANN process simulation results (training data set) when only an initial state is supplied and 132 hours of process operation time are simulated. Experimental data points are averaged for convenience, and the measurement errors (error bars) are not presented in the figures. The MAPE (mostly below 5%) indicates the high accuracy of the current ANN. The slightly higher MAPE in nitrogen at the beginning (12 hours to 36 hours) is explained at Section 3.3. Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.

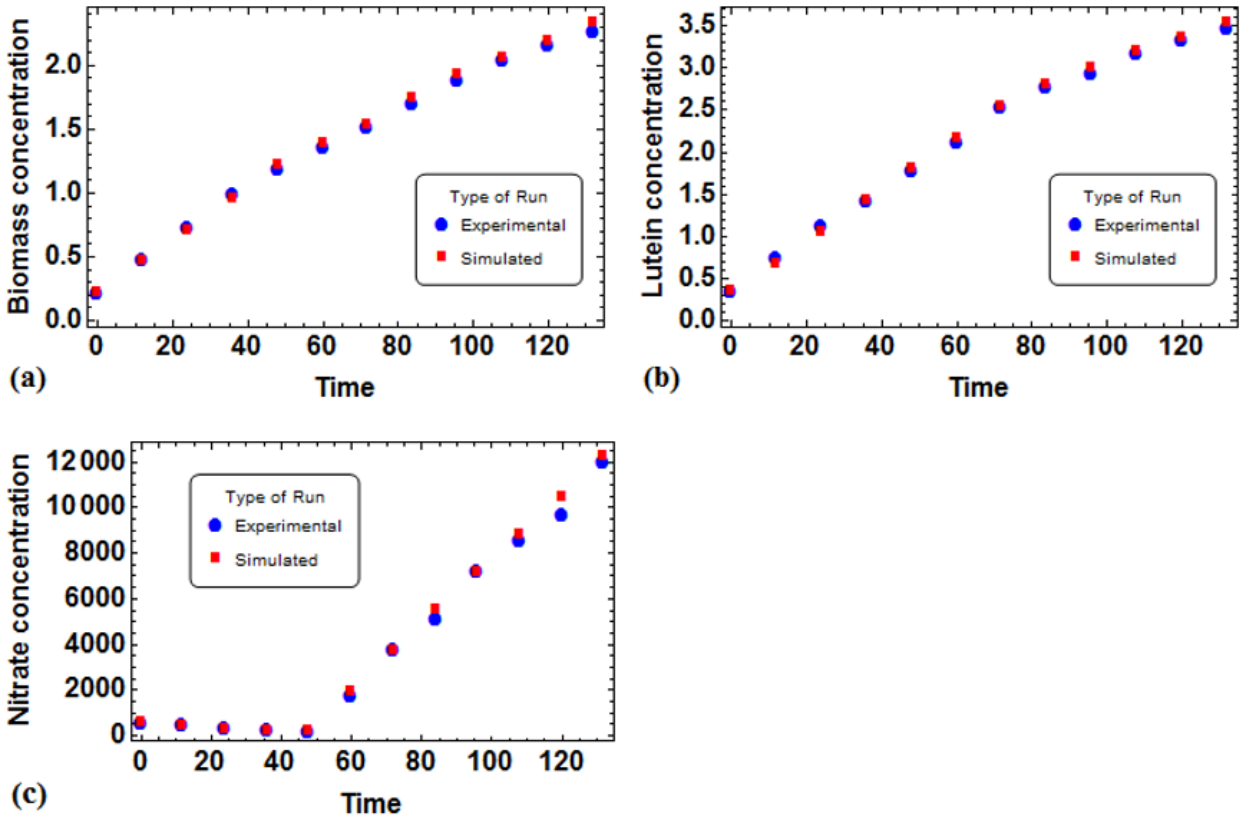


Figure 5: Two hidden layers ANN process simulation results (training data set) when only an initial state is supplied and 132 hours of process operation time are simulated. The MAPE of this ANN is mainly below 3%. Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.

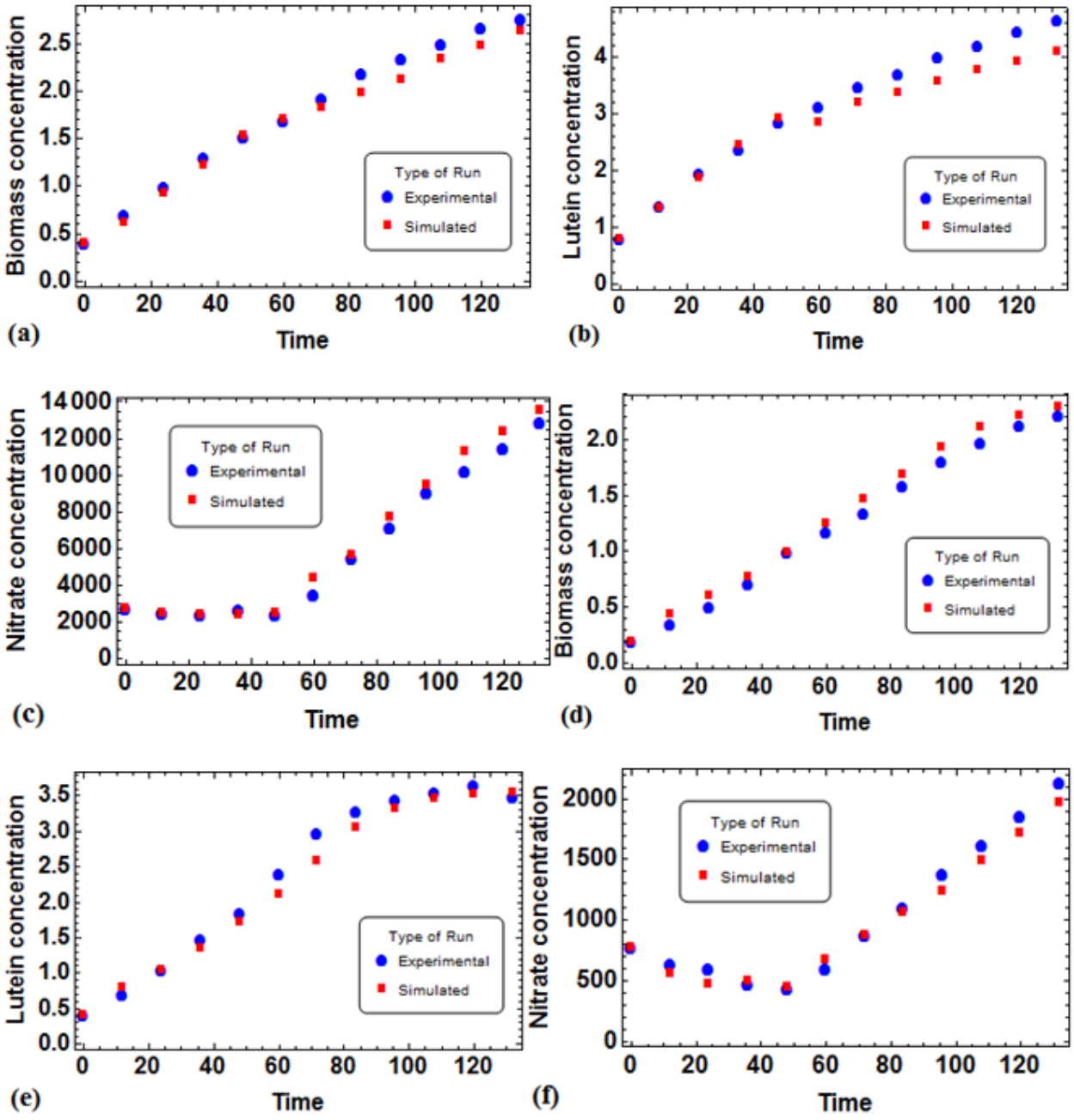


Figure 6: Comparison of the two hidden layers ANN real-time prediction results with real experimental data. (a), (b), and (c): Experiment Test 1. (d), (e), and (f): Experiment Test 2. Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.

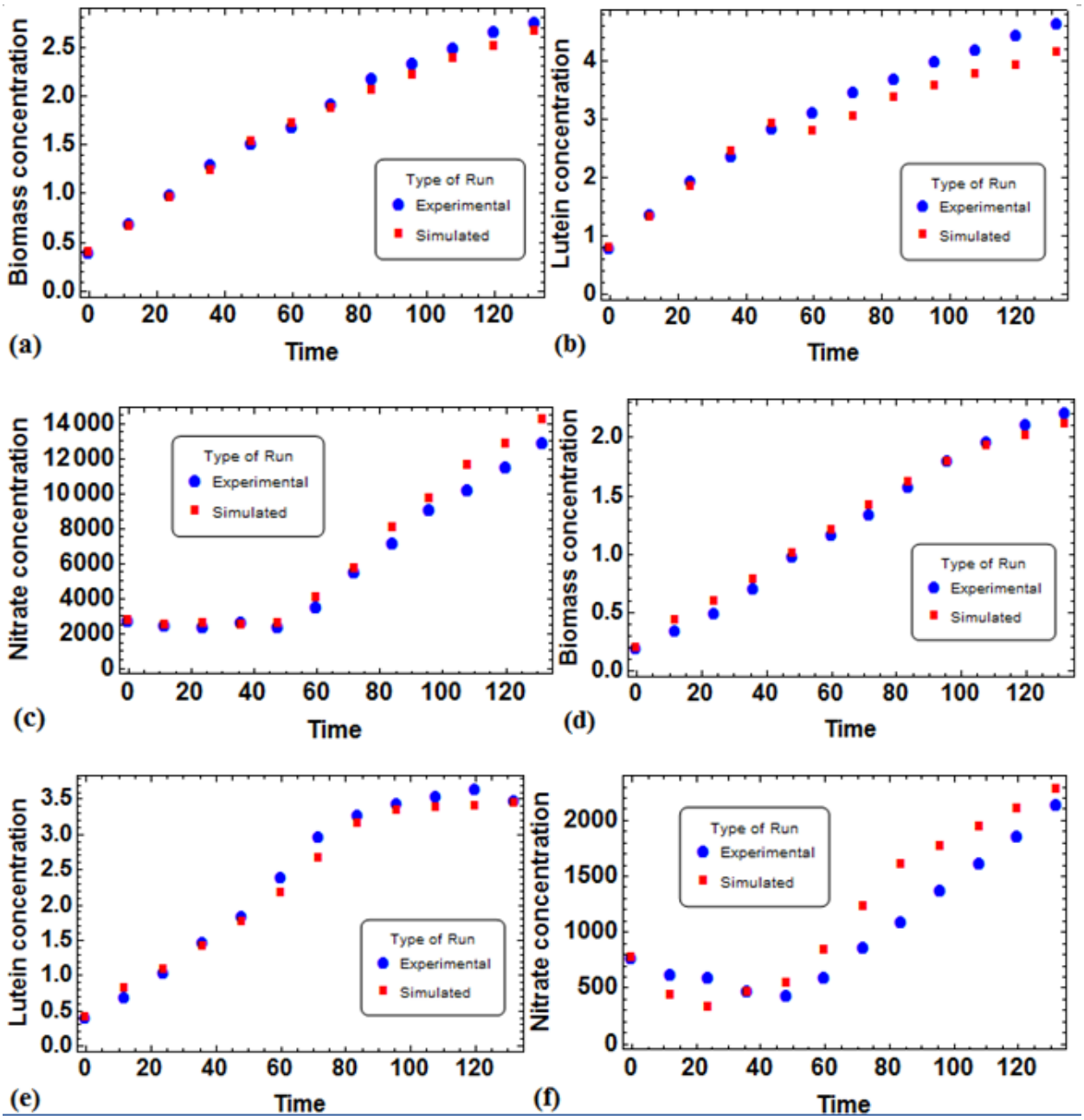


Figure 7: Comparison of the one hidden layer ANN real-time prediction results with real experimental data. (a), (b), and (c): Experiment Test 1. (d), (e), and (f): Experiment Test 2. Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.

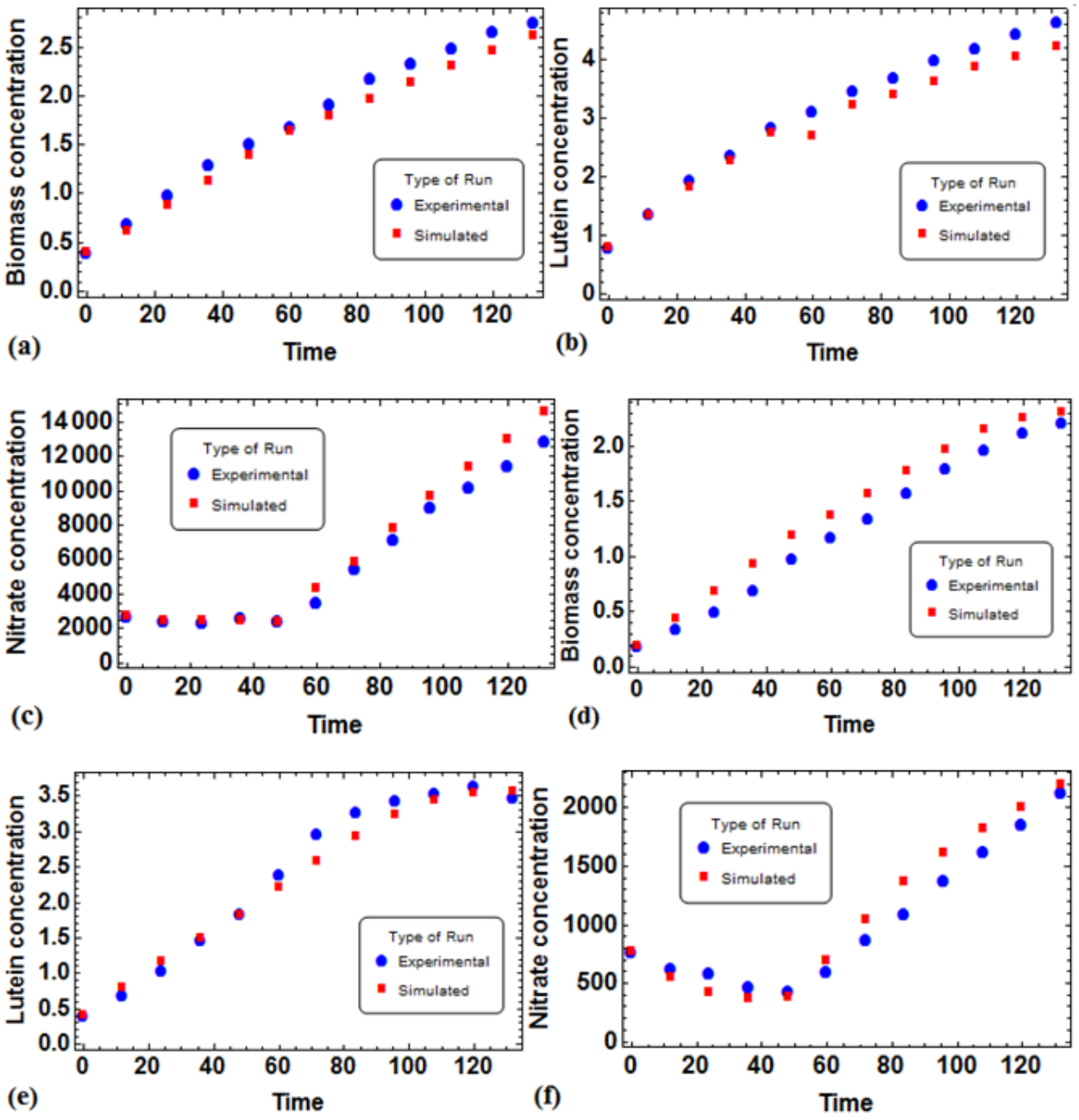


Figure 8: Comparison of the two hidden layers ANN offline prediction results with real experimental data. (a), (b), and (c): Experiment Test 1. (d), (e), and (f): Experiment Test 2. Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.

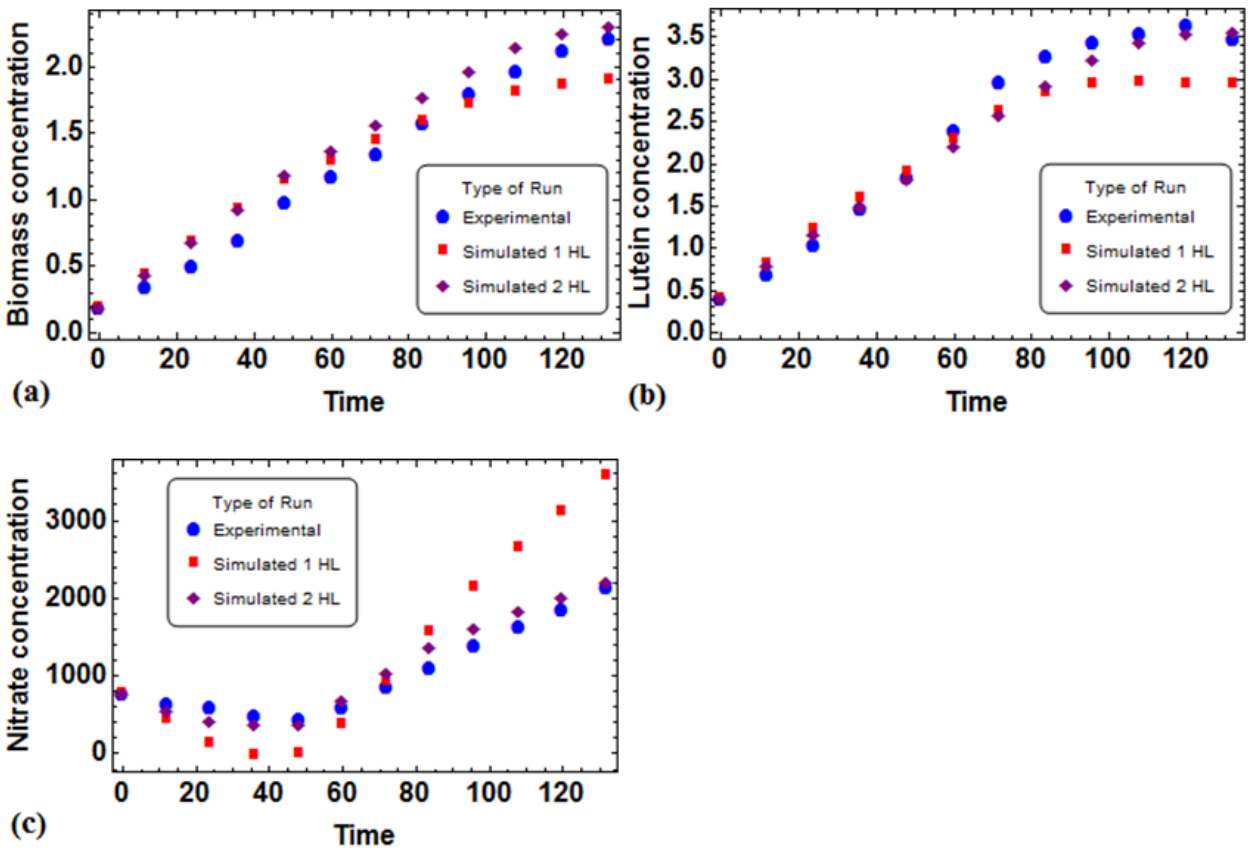
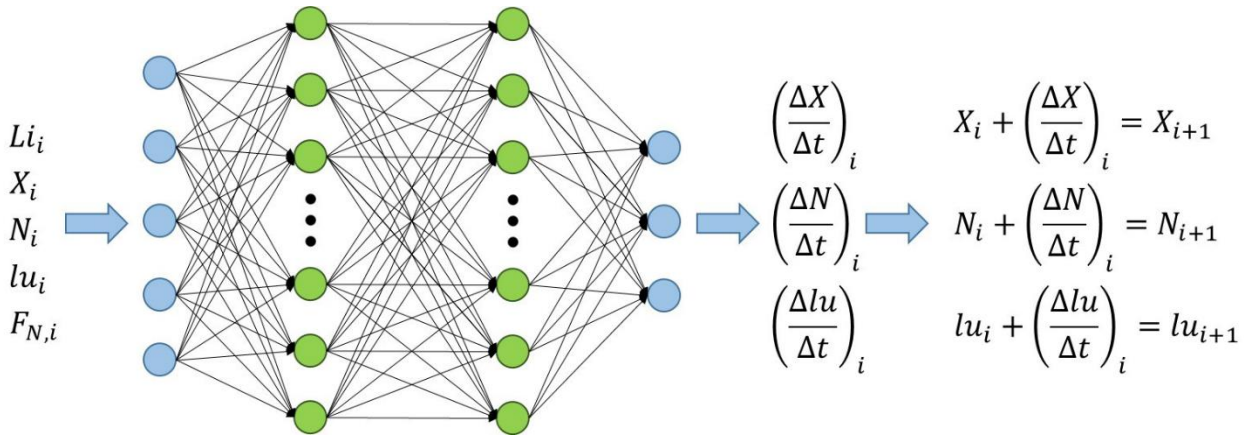


Figure 9: Comparison of prediction results between one hidden layer ANN and two hidden layers ANN in the offline framework (Experiment Test 2). Biomass concentration: g L^{-1}, lutein concentration and nitrate concentration: mg L^{-1}.



Graphical Table of Contents: Two robust artificial neural networks were constructed to simulate the dynamic behaviour of microalgae growth and lutein production; different advanced strategies were incorporated to guarantee the accuracy of the constructed models, including determining the optimal network structure through a hyper-parameter selection framework, generating artificial data sets by embedding appropriate random noise, and rescaling model inputs through standardisation; the accuracy and predictive power of the models for long-term dynamic bioprocess simulation in real-time and offline frameworks were demonstrated and verified experimentally.