Dongda Zhang ORCID iD: 0000-0001-5956-4618

Review of advanced physical and data-driven models for dynamic bioprocess simulation: Case study of algae-bacteria consortium wastewater treatment

Ehecatl Antonio Del Rio-Chanona[1,ǂ], Xiaoyan Cong[2,ǂ], Eric Bradford[3], Dongda Zhang[1,4,*], Keju Jing[2,*]

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

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

3: Engineering Cybernetics, Norwegian University of Science and Technology, Trondheim, Norway.

4: Centre for Process Integration, University of Manchester, Oxford Road, Manchester, M1 3BU, UK.

ǂ: These authors contributed equally to this work.

*: Corresponding authors, email: dongda.zhang@manchester.ac.uk, tel: 44 (0)161 306 5153 (Dongda Zhang); jkj@xmu.edu.cn, tel: 86 592 2186038 (Keju Jing).

Running title: Machine learning for bioprocess modelling

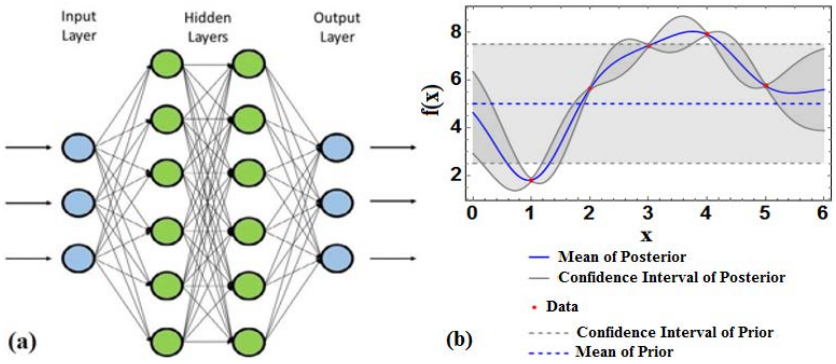
This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/bit.26881.

Abstract

Microorganism production and remediation processes are of critical importance to the next generation of sustainable industries. Undertaking mathematical treatment of dynamic biosystems operating at any spatial or temporal scale is essential to guarantee their performance and safety. However, constructing physical models remains a challenge due to the extreme complexity of process biological mechanisms. Data-driven models also encounter severe limitations because datasets from large-scale bioprocesses are often scarce without complete information and on a restricted operational space. To fill this gap, the current research compares the performance of advanced physical and data-driven models for dynamic bioprocess simulations subject to incomplete and scarce datasets, which to the best of our knowledge has never been addressed before. In specific, kinetic models were constructed by integrating different classic models, and state-of-the-art hyperparameter selection frameworks were developed to design artificial neural networks and Gaussian process regression models. An algae-bacteria consortium wastewater treatment process was selected to test the accuracy of these modelling strategies, as it is one of the most sophisticated biosystems due to the intricate mutualistic and competitive interactions. Based on the current results and available data, a heuristic model selection procedure is provided. This research paves the way to facilitate future bioprocess modelling.

Graphical Abstract

Microorganism production and remediation processes are of critical importance to the next generation of sustainable industries. Undertaking mathematical treatment of dynamic biosystems operating at any spatial or temporal scale is essential to guarantee their performance and safety. However, constructing physical models remains a challenge due to the extreme complexity of process biological mechanisms.



Keywords: Gaussian processes, artificial neural network, kinetic modelling, scarce dataset, algae-bacteria consortium

Introduction

Microorganism based industrial biotechnologies have drawn great attention within the last decade due to their applications in sustainable production and environmental remediation. Microalgae have been extensively studied to produce a variety of renewables *e.g.* biohydrogen, transportation fuels, food supplements and high-value bioproducts (Dongda Zhang and Vassiliadis 2015; Harun et al. 2018; Jiao et al. 2017). They can utilise solar energy and CO_2 for bioproduct synthesis without the necessity of occupying arable land and competing with agricultural plants. Several algal products have been commercialised in the US, China, and the Middle East, and their global market has been predicted to reach over $5.1 billion by 2023 (Wood 2018).

Traditional fermentation processes also play a vital role for industrial scale production of a broad range of commodities including bulk chemicals, polymers, pharmaceuticals, and food additives (Jing et al. 2018; Wang et al. 2015; Bankar et al. 2014). Their global market demand has been expected to reach over $2.4 trillion in 2025 (John 2018). Meanwhile, algae-bacteria consortia have also been used for wastewater treatment and detoxification of environmental pollutants (Jia and Yuan 2016). They have been reported to effectively remove different nitrogen, phosphorus and carbon source (Delgadillo-Mirquez et al. 2016; He et al. 2013).

Bioproduction and bioremediation processes are conducted dynamically in a batch or fed-batch operation. To improve their efficiency and safety, it is vital to allow the mathematical treatment of bioprocesses to improve performance and reliability through advanced optimisation and control methods. As a result, a rigorous model capable of simulating complex biological dynamics is essential. Conventionally, this was achieved by constructing physical models based on biological mechanisms. Kinetic models, a class of grey-box models, are principally used for bioprocess modelling, optimisation, control, and design (Quinn et al. 2011; del Rio-Chanona, et al. 2017). Kinetic models lump the large number of metabolic pathways into a small set of differential equations to model cell growth, substrates uptake, and product production. Classic kinetic models for fermentation include the Monod model, the Droop model, the Contois model, and the Luedeking–Piret model, each one designed by distinct assumptions and used under different circumstances (Vatcheva et al. 2006). For algal systems, kinetic models that include light effects have also been designed (Quinn et al. 2011; D. Zhang et al. 2015). However, for multi-strain co-culturing systems (*e.g.* algae-bacteria consortium), measuring cell growth and nutrients uptake of each strain is difficult, causing a challenge for process modelling.

Datasets in many bioprocesses are scarce and involve time-series with high uncertainty. They are usually incomplete, meaning that part of the information is missing due to equipment limitation or labour shortage. Kinetic models can handle these issues effectively, but their application has been severely limited due to the very high complexity of mechanisms underlying the biosystems. For example, photo-production or algae-bacteria consortium remediation processes are affected by various factors including multiple nutrients, light, and temperature. Intricate interactions between these factors are poorly understood, making it challenging to construct an accurate model. Thus, kinetic model parameters are often assigned different values to model the behaviour of biological processes well for a specific range of operating conditions in a particular experiment (Adesanya et al. 2014; He et al. 2013). This causes the loss of predictive power of kinetic models, as they then cannot predict well the biological process at conditions distinct from those used in the experiments. Hence, machine learning (ML) methods have been increasingly applied as an alternative.

Artificial neural networks (ANN) are one of the earliest ML methods used in chemical engineering (He et al. 2013). Being black-box models, they can estimate complicated relationships between inputs and outputs without the necessity of understanding the detailed physical mechanisms. They have been utilised to model and optimise microbial bioproduction processes, yielding substantial increases (85% to 187%) on productivity of biorenewables (del Rio-Chanona et al. 2016; Dineshkumar et al. 2015). Recently, there is an emerging effort to exploit Gaussian process (GP) regression, a cutting-edge ML method, for bioprocess modelling, optimisation, and monitoring (Bradford et al. 2018; Tulsyan et al. 2018). GP regression provides predictions as Gaussian distributed variables conditioned on the

available data. They possess an excellent feature compared to most ML and physical models, which is to predict output uncertainty. This is especially important to biosystems due to their high uncertainty arising from the sophisticated and sensitive metabolisms. Despite successful applications in other fields, ML methods have encountered critical bottlenecks in bioprocesses due to the small size and incompleteness of the datasets. As they are data-driven models, collecting large datasets is vital for their construction. Meanwhile, having a full record of measurements at each time step is essential for them to learn system dynamics. Nonetheless, neither of these pre-requisites can be easily satisfied for biosystems.

This study aims to compare performance (*i.e.* simulation accuracy, predictive capability) of different types of models when confronting applications with scarce datasets, thus providing suggestions for future modelling studies. Algae-bacteria consortium wastewater treatment is selected as the case study due to its high complexity. State-of-the-art model construction strategies were adopted with their advantages and disadvantages thoroughly discussed. The most reliable models were then used to improve understandings of the underlying system.

2. Material and methods

2.1 Strains selection and medium

Alga *Chlorella vulgaris* GY-H4 was purchased from the Institute of Hydrobiology (IHB), Chinese Academy of Sciences, China; and bacterium *Bacillus subtilis* was obtained from earlier work in our laboratory and stored at the Culture Collection of Xiamen University. Prior to the experiments in synthetic wastewater (SWW), algal and bacterial cells were pre-cultured in the BG-11 medium and the Luria-Bertani (LB) medium, respectively. *C .vulgaris* and *B. subtilis* were inoculated separately in both

high and low concentration SWW mediums. The high concentration SWW was initially composed of (per L of distilled water): 500 mg Glucose, 1750 mg $NaHCO_3$, 727 mg $NaNO_3$, 83.3 mg KH_2PO_4, 7 mg NaCl, 4 mg $CaCl_2 \cdot 2H_2O$, 75 mg $MgSO_4 \cdot 7H_2O$, 2.5 mg $FeSO_4$, 20 mg EDTA, 0.00125 mg $ZnSO_4$, 0.0025 mg $MnSO_4$, 0.0125 mg H_3BO_3, 0.0125 mg $Co(NO_3)_2$, 0.0125 mg Na_2MoO_4, and 6.25×10^6 mg $CuSO_4$. This resulted in 200 mg/L dissolved organic carbon (DOC), 120 mg/L N-NO_3^-, and 19 mg/L TP-PO_4^{3-}. The low concentration SWW contains (per L of distilled water): 100 mg Glucose, 350 mg $NaHCO_3$, 115 mg $NaNO_3$, 13.2 mg KH_2PO_4, 7 mg NaCl, 4 mg, $CaCl_2 \cdot 2H_2O$, 75 mg $MgSO_4 \cdot 7H_2O$, 2.5 mg $FeSO_4$, 20 mg EDTA, 0.00125 mg $ZnSO_4$, 0.0025 mg $MnSO_4$, 0.0125 mg H_3BO_3, 0.0125 mg $Co(NO_3)_2$, 0.0125 mg Na_2MoO_4, and 6.25×10^6 mg $CuSO_4$. This resulted in 40mg/L DOC, 19 mg/L N-NO_3^-, and 3 mg/L TP-PO_4^{3-}.

2.2 Culture methods and experiment setup

Bacterial experiments were conducted in a 500mL baffled flask containing 100 mL SWW medium and cultivated at 28°C, 200 rpm for 8 days, with an initial inoculum size of 0.24 g/L. The algal and algae-bacteria consortium experiments were conducted in a 1L photobioreactor (PBR) equipped with an external light source mounted on both sides. Light intensity was 300 μmol/m^2/s and aeration rate was 0.1 vvm with 2.5% CO_2. Initial culture volume was 800 mL SSW medium and the cultures were incubated for 8 days at 25-28°C. Initial biomass concentration for the algal experiments was 0.24 g/L. In the consortium experiments, the same inoculum size of algae and bacteria was added into the PBR with a joint concentration of 0.48 g/L. The consortium was also cultivated in the sterilized SWW with high and low concentrations of glucose (500 and 100mg/L), TN-NO_3^-(120 and 19 mg/L) and TP-

PO_4^{3-} (19 and 3 mg/L), respectively. The culture pH was maintained at 7 to 8. Liquid samples were collected from the culture broth at set time intervals to measure cell concentration, DOC, TP and TN. Experiments were conducted in triplicate and are summarised in Table I.

2.3 Analytical procedures

Biomass concentration was measured through optical density at a wavelength of 680 nm (OD_{680}) and recorded as dry weight (g/L). Biomass was harvested by centrifugation (5000 rpm, 5 min) and washed three times using reverse osmosis treated water. During the experiments, carbon concentration was determined by a TOC analyser (LiquiTOC II, Elementar, Germany) from filtrated samples (0.45 μm). NO_3^- and PO_4^{3-} ions from the filtrated (0.20 μm) wastewater was analysed by an Ion Chromatograph (ICS-5000, Dionex, Italy).

3. Modelling methodology

3.1 Dataset augmentation for the construction of data-driven models

Datasets from the four single strain processes (Table I) were used for model construction. Data points were measured once every 6 hours, some of which were excluded to resemble industrial cases. For kinetic models, the datasets were used directly for parameter estimation. For ML models, two strategies were applied with their advantages discussed in Section 4. The first is to fill missing information by linearly interpolating existing data, and the second is to generate a set of artificial datasets by embedding adequate noise (±3% standard deviation given the equipment precision) into the original datasets (del Rio-Chanona, et al. 2017). Then, the augmented datasets were normalised to train ML models.

3.2 Construction of kinetic models

A number of kinetic models were adopted and modified in this study. As each model parameter has a unique physical meaning, their number in a kinetic model is less than that in a ML model. The model structure which represents best the dynamics of bacterial experiments is shown as Eqs. 1(a)-1(d), built on the original Monod model, Logistic model, and Luedeking–Piret model (Zhang et al. 2015). The first term on the right-hand-side (RHS) in Eq. 1(a) represents cell growth, with the second calculating cell decay. The first term on the RHS in Eqs. 1(b)-1(d) denotes cell-growth dependent uptake of each substrate, with the second term estimating cell-growth independent consumptions (*e.g.* used for cell maintenance).

$$\frac{dX}{dt} = \mu \cdot \frac{N}{N+K_N} \cdot \frac{C}{C+K_C} \cdot \frac{P}{P+K_P} \cdot X - \mu_d \cdot X^2 \; 1(a)$$

$$\frac{dC}{dt} = -Y_{C1} \cdot \left(\mu \cdot \frac{N}{N+K_N} \cdot \frac{C}{C+K_C} \cdot \frac{P}{P+K_P} \cdot X - \mu_d \cdot X^2\right) - Y_{C2} \cdot X \; 1(b)$$

$$\frac{dN}{dt} = -Y_{N1} \cdot \left(\mu \cdot \frac{N}{N+K_N} \cdot \frac{C}{C+K_C} \cdot \frac{P}{P+K_P} \cdot X - \mu_d \cdot X^2\right) - Y_{N2} \cdot X \; 1(c)$$

$$\frac{dP}{dt} = -Y_{P1} \cdot \left(\mu \cdot \frac{N}{N+K_N} \cdot \frac{C}{C+K_C} \cdot \frac{P}{P+K_P} \cdot X - \mu_d \cdot X^2\right) - Y_{P2} \cdot X \; 1(d)$$

where X, N, C, P are concentrations of biomass, nitrate, glucose, and phosphate, respectively; K_i is half-velocity coefficient of substrate i; Y_{i1} and Y_{i2} are growth-dependent and growth-independent yield coefficient of i; μ and μ_d are specific growth and decay rate.

The best kinetic model structure for algal processes (also adopted from the three classical models) is shown in Eqs. 2(a)-2(d), and all terms on the RHS denote the

same meaning as those in Eqs. 1(a)-1(d). As light intensity was fixed, to avoid parameter identifiability and over-fitting issues, its effects are grouped into the specific growth rate term and not listed separately.

$$\frac{dX}{dt} = \mu \cdot \frac{N}{N + K_N} \cdot \frac{C}{C + K_C} \cdot \frac{P}{P + K_P} \cdot X \cdot \left(1 - \frac{X}{X_{max}}\right) \quad 2(a)$$

$$\frac{dC}{dt} = -Y_{C1} \cdot \mu \cdot \frac{N}{N + K_N} \cdot \frac{C}{C + K_C} \cdot \frac{P}{P + K_P} \cdot X \cdot \left(1 - \frac{X}{X_{max}}\right) - Y_{C2} \cdot X \quad 2(b)$$

$$\frac{dN}{dt} = -Y_{N1} \cdot \mu \cdot \frac{N}{N + K_N} \cdot \frac{C}{C + K_C} \cdot \frac{P}{P + K_P} \cdot X \cdot \left(1 - \frac{X}{X_{max}}\right) - Y_{N2} \cdot X \quad 2(c)$$

$$\frac{dP}{dt} = -Y_{P1} \cdot \mu \cdot \frac{N}{N + K_N} \cdot \frac{C}{C + K_C} \cdot \frac{P}{P + K_P} \cdot X \cdot \left(1 - \frac{X}{X_{max}}\right) - Y_{P2} \cdot X \quad 2(d)$$

where X_{max} denotes the maximum biomass concentration.

Parameter estimation was conducted by a weighted nonlinear least squares optimisation problem. Given the high nonlinearity and stiffness, the differential equations were discretised by orthogonal collocation over finite elements in time using Radau roots (del Rio-Chanona et al. 2015; Kameswaran and Biegler 2008). The problem was solved using the interior point nonlinear optimisation solver IPOPT through a multi-start framework in the parameter space (Wächter and Biegler 2006). This was programmed in the Python optimisation environment Pyomo (Hart et al. 2012). The models were simulated in Mathematica 11.

3.3 Construction of machine learning based models

3.3.1 Construction of Artificial Neural Networks

An ANN (Fig. 1(a)) comprises an input and an output layer, and several hidden layers, each of which contains several neurons to store activation functions (*e.g.* sigmoid function, Eq. 3) and formulate relations between inputs and outputs. To apply to a dynamic system, an ANN is designed by feeding the system's current states to predict future ones at the next time step. By recursively using the ANN, behaviour of a process over the entire course can be modelled (del Rio-Chanona et al. 2016). Another approach is to use Recurrent Neural Networks which is structured specifically to model time-series events (Valdez-Castro, Baruch, and Barrera-Cortés 2003). This work builds on the feedforward ANN without loss of generality.

$$y_j = \frac{1}{1 + \exp\left(-\left(\sum_i x_i \cdot w_{ij} + b_j\right)\right)} \quad (3)$$

where y_j is the output from neuron j, x_i is the input from the i^{th} neuron in the previous layer, w_{ij} is the weight of x_i, and b_j is the bias.

To construct an accurate ANN, both parameters and hyperparameters must be optimised. Parameter optimisation (weights and bias) follows the standard backpropagation method. The key to obtain a rigorous ANN lies in the estimation of hyperparameters (numbers of neurons, layers, and training epochs). Increasing these numbers increases the model complexity, which gives a better fit of the training data but increases the risk of over-fitting, worsening the predictive power for data outside of the training data. Higher model complexity also leads to higher computational costs. In our previous work (del Rio-Chanona et al. 2017), a hyperparameter selection

framework, namely “elbow rule”, was adopted to balance the trade-off between model accuracy, computational cost and over-fitting. This strategy was refined by examining the optimal size of artificial datasets in this work. Another technique is the *k*-fold method, where a selection of *N-1* from the *N* datasets is used for ANN training and the remaining one is used to estimate the maximum prediction error of this ANN. Then, another *N-1* subsets are selected to repeat this procedure until the best model is identified.

The *k*-fold method is applied when the size of datasets is greater than 3. This is not the current case as each system only has two experiments governed by different kinetics (Table I). Hence, the two datasets must be fitted together. As a result, 70% of data points from both sets were randomly chosen to train ANN and the rest was used for cross validation. Inputs of the current ANN includes concentrations of biomass and all nutrients (*i.e.* 4 inputs), with outputs being changes of these state variables after 6 hours. Two ANNs were constructed, one for algae and the other for bacteria. Through the refined “elbow rule”, optimal structure of both ANNs was found to contain 2 hidden layers, each including 8 neurons. 100 artificial datasets (zero-mean Gaussian noise with 3% standard deviation) were generated for the algal ANN, and 50 for the bacterial ANN. Number of training epochs was 5,000 for the algal ANN and 2,000 for the bacterial ANN. The larger number of artificial datasets and training epochs required for the algal ANN construction may indicate that the algal process involves more complex metabolic mechanisms compared to the bacterial process. Once the optimal structure identified, the ANNs were trained again using all available datasets to complete model construction. All these implementations were carried out in Mathematica 11.

3.3.2 Construction of Gaussian Processes

In this section a brief description of GPs is given, for more information please refer to (Rasmussen & Williams 2006), we provide a detailed explanation of GPs applied to bioprocesses in (Bradford et al. 2018). Unlike ANNs, GPs provide an uncertainty measure representing the prediction uncertainty of the unknown function given the availability of only limited amounts of data. This uncertainty can be used to evaluate the reliability of GP predictions to prevent over-optimistic conclusions. GP regression aims to model a latent function $f(\mathbf{x})$ given noisy measurements. The relationship between the function $f(\mathbf{x})$ and the measurements can be expressed as follows (Kirk and Stumpf 2009):

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (4)$$

where $y(\mathbf{x})$ denotes the measurement of $f(\mathbf{x})$ at $\mathbf{x}$ and ε the corresponding measurement noise assumed to follow a normal distribution with zero mean and variance σ^2.

The GP regression starts with the definition of a *prior* GP distribution (Fig. 1(b)), which describes the function to be modelled $f(\mathbf{x})$ before any data is used and hence encapsulates the assumptions made on this function *e.g.* continuity or smoothness. The prior takes the form:

$$f(\mathbf{x}) \sim GP\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \quad (5)$$

where $k(\mathbf{x}, \mathbf{x}')$ is the covariance function, $\mathbf{x}, \mathbf{x}'$ are arbitrary inputs, and $m(\mathbf{x})$ is the mean function.

Now assume we are given a set of p observations $y_1, \ldots, y_p$ of $f(\mathbf{x})$ evaluated at different inputs $\mathbf{x}_1, \ldots, \mathbf{x}_p$. The *prior* can then be updated using this data to obtain the *posterior* GP distribution (Fig. 1(c)), which is the updated distribution of $f(\mathbf{x})$ given the available data. The posterior GP can subsequently be employed to estimate the conditional probability of $f(\mathbf{x}_*)$ represented by a Gaussian distribution at an arbitrary input $\mathbf{x}_*$ given the observed information. The mean in this context represents the prediction of $f(\mathbf{x}_*)$, whilst the variance denotes the uncertainty. As GPs are non-parametric methods, their accuracy heavily depends on the selection of their hyperparameters (parameters of the mean function $m(\mathbf{x})$, covariance function $k(\mathbf{x}, \mathbf{x}')$ and measurement noise ε). When encountering applications with scarce datasets, the maximum a posteriori method is recommended to optimise GPs' hyperparameters rather than the maximum likelihood method given its advantage in preventing over-fitting (Rasmussen and Williams 2006). So far, we have described multi-input, single output usage of GPs, however we are interested in the multi-input, multi-output case. This can be achieved by training separate GPs for each output, which are then used together for multi-input, multi-output predictions.

In this study, 4 GPs were constructed to simulate algal processes (one for each state), the input of each of which is the concentration of biomass and nutrients, with the output of each referring to the change of concentration of one specific state variable after 6 hours. The same GP framework was also designed for the bacteria system. 3 artificial datasets were generated to train the GPs in each case. Construction of the GPs was programmed in Mathematica 11. It is worth mentioning that in principle it is feasible to construct four separate ANNs. However, in practice ANN is often designed as a multi-input, multi-output (MIMO) model because it requires less computational

time for model training and meanwhile able to achieve the same accuracy compared to several multi-input, single-output (MISO) ANN models.

4. Results and Discussion

4.1 Comparison of physical models and machine learning based models

Once constructed, all models were initially used to simulate the 4 single strain processes from which their parameters were estimated. To test accuracy, the *offline modelling framework* was used in which only initial conditions were given and the models must simulate over the entire process time course of operation. This is accomplished by recursively using the fitted models, *i.e.* by using the initial condition to predict one-step ahead and then using this prediction as the next input to obtain the next prediction. It should be noted that for each strain, the two experimental datasets were fed together to train one ANN and one GP framework; whilst they were used separately to estimate two sets of parameter values for kinetic models (Table II). A discussion of this implementation is shown in Section 4.1.3.

4.1.1 Comparison on algae wastewater treatment process

Figs. 2-3 show the model simulation results of algae processes. It can be seen that in the high concentration experiment the kinetic model has larger errors than the machine learning based models, particularly when simulating nitrate (Fig. 2(b)) and glucose (Fig. 2(c)), indicating a mismatch between model structure and biological mechanisms; whilst all models act similarly when simulating biomass growth and phosphate uptake. The kinetic model fits well in the low concentration experiment, with mild errors (Fig. 3(d) slightly larger than the other methods) when modelling phosphate consumption. In terms of the two data-driven models, in most cases there is

no distinguishable difference between their simulation results. This is expected, since ANNs and GPs will always be able to fit noiseless training data exactly. This is further highlighted by the GP having very low uncertainty (not visible). Nonetheless, the ANN overestimates the final nitrate concentration consistently (Fig. 3(b)), unlike the GP which has a high prediction quality throughout and can hence be regarded as more reliable.

4.1.2 Comparison on bacteria wastewater treatment process

Figs. 4-5 show the model simulation results of bacteria processes. Opposite to the algal system, it is observed that the kinetic model can represent the two bacterial experiments well for the most part, except for the overlook of final nitrate uptake in the high concentration experiment (Fig. 4(b)) and glucose uptake in the low concentration experiment (Fig. 5(c)). The ANN, however, overestimates concentrations of nitrate and phosphate in later stage of the low concentration process (Figs. 5(b) and 5(d)), although this is not significant. It is worth stressing that some data points in these experiments have large measurement and stochastic noise as they deviate from the system's dynamic trajectory (*e.g.* nitrate at the 36^{th} hour in Fig. 4(b) and 24^{th} hour in Fig. 5(b), biomass at the 24^{th} hour in Fig. 5(a), glucose at the 36^{th} hour in Fig. 5(c)). The kinetic model can take advantage of its structure to filter out the noise as it is constructed by biochemical mechanisms; whilst neither of the data-driven models is able to remove the noise since they assume input of the training data does not have stochastic error.

4.1.3 Discussion on simulation of complex bioprocesses

This study shows that compared to the data-driven models, the kinetic model is successful in representing the bacterial processes, but large deviations are found when simulating algal systems, indicating its inadequacy for process predictions and exploration of algae-bacteria interactions. Thus, ANNs and GPs are used in the next section. However, a comprehensive comparison between kinetic models and machine learning models is conducted here.

From the time-efficiency aspect, constructing a kinetic model is in general considerably more time consuming (summarised in Table III). For instance, over 15 structures of kinetic models were designed in this study by adopting and amending a number of advanced models with various biological hypotheses. However, growth of cells and uptake of nutrients are subject to distinct mechanisms under the two extreme conditions, making it infeasible to obtain a single structure or set of parameter values that describe both mechanisms well. The current work successfully identified the optimal model structure valid in algae and bacteria systems for the two extreme conditions, and all parameters have a valid physical interpretation. Nonetheless, when gathering both datasets to estimate parameter values, the model fails to represent either of them as the parameters were calculated to compromise the contradictory behaviours. Thus, each dataset was used to estimate its own parameters so that the model can represent the different mechanisms. This unavoidably sacrifices the model prediction ability. In contrast, the key to train a machine learning model is to identify the optimal hyperparameters. Specific to ANNs and GPs, effective hyperparameter selection frameworks have been proposed in our studies and refined in this work.

Therefore, designing a data-driven model only took a few days whilst that for a kinetic model cost several weeks (Table III).

From the datasets perspective, ANNs require many large datasets. This was solved by generating artificial datasets. As dynamics of fermentation and photo-production processes do not change drastically in general, it is acceptable to fill the missing data by linear interpolation over a short time span. For GPs, the number of artificial datasets must be selected cautiously. Adding artificial datasets can consolidate GPs' accuracy in predicting the mean of the output. But they will also shape GPs' posterior distribution and interfere with the GPs' prediction on the output uncertainty, thus deteriorating GPs' performance in robust optimisation. As a result, GPs require much less artificial datasets (*e.g.* 3 sets in this study) than ANNs (*e.g.* 50-100 sets in this study). Hypothetically, artificial datasets for GPs can be substituted by carefully tuning the corresponding measurement noise term; we however leave this matter to be addressed by future research. In contrast, a kinetic model does not need complete or large datasets, and their parameter estimation method can nullify the use of artificial datasets. In fact, adding artificial datasets may be detrimental to kinetic models as it amplifies the scale and complexity of the parameter estimation problem. This should be avoided if a kinetic model is highly nonlinear and stiff. It is important to stress that kinetic models are vital in many applications such as process scale-up and bioreactor design which cannot be replaced by machine learning methods. Thus, the conclusion that kinetic models are less efficient cannot be generalised.

Finally, it should be observed that an important factor – light intensity – is not included in the current kinetic model for algal process simulation. This is because incident light intensity in the current experiments was fixed constant, thus it is

difficult to accurately identify values of relevant kinetic parameters. Hence, future experiments should be implemented with different light intensities. It is expected that through the inclusion of light intensity effects, the kinetic model may present a better simulation performance. However, adding more parameters will complicate the kinetic model structure; this trade-off should be balanced in future research.

4.2 Process modelling and mechanism exploration on algae-bacteria interactions

To investigate algae-bacteria interactions during wastewater treatment, three extreme cases: algae completely inhibiting bacteria growth (Case 1), bacteria completely inhibiting algae growth (Case 2), and algae and bacteria growing independently (Case 3) were simulated and compared to the experimental data. In the first two cases, the consortium process is reduced to a single strain system and the offline modelling framework was adopted. In Case 3, uptake of nutrients was assumed to be the sum of that consumed by algae (predicted by algal models) and that by bacteria (predicted by bacterial models). Strictly speaking, this approximation only holds within a small time interval. Hence, in Case 3 the *online modelling framework* was used such that models only predict nutrient concentrations one step ahead and then experimental data at the next time step are fed as model input for further predictions. It is noted that individual biomass concentrations cannot be measured, thus in all cases algal and bacterial concentrations were predicted through the offline framework.

4.2.1 Investigation of algae-bacteria interaction under high nutrient concentrations

From Fig. 6, it is seen that ANNs and GPs predict similar results in Cases 1 and 2 (Figs. 6(a), (c), (e), (g)), except for the final nitrate concentration in Fig. 6(c). The GPs predict a closer result to the data compared to the ANNs in Case 3 (Figs. 6(b), (d),

(f), (h)), indicating their better predictive capability. Large deviation exhibited in the ANNs (*e.g.* Fig. 6(b)) may be attributed to the propagated errors for prediction of algae and bacteria concentrations through the offline framework. Hence, the GPs' results are chosen for further analysis.

From the figures, it is firstly concluded that algae growth is noticeable. This is obtained by comparing the offline prediction results with the experimental data. Figs. 6(c) and 6(e) show that nitrate and phosphate are barely consumed in Case 2 (bacteria growth dominates) but rapidly decreased in Case 1 (algae growth dominates). This is consistent with previous studies in which algae instead of bacteria are found to mainly consume nitrogen and phosphorus (Hernandez et al. 2009; Liang et al. 2013). Secondly, there exists a mild algae-bacteria competition for organic carbon. This is because glucose concentrations in Cases 1 and 2 are similar to the data (Fig. 6(g)) and final algae cell concentration in Case 1 is almost the same as the experimental result (Fig. 6(a)). Thus, if there is no competition, total biomass concentration of the consortium should be higher than the experimental data with glucose being lower. Given that Case 3 (independent growth of algae and bacteria) also predicts similar dynamics to the data (Figs. 6(b), (f), (h)), it is suggested that this competition should not be serious and may not be the primary interaction.

Most importantly, it is seen that nitrate uptake in Cases 1 and 2 and even the sum of them are markedly slower than the real observation (Fig. 6(c)). However, Case 3 is very similar to the data (Fig. 6(d)), indicating that the presence of both strains may significantly accelerate nitrate consumption. This has been reported by several research that bacteria can promote algal nitrate uptake (Hernandez et al. 2009; Subashchandrabose et al. 2011). A previous work using similar bacteria and algae

species to the current study claims that 78% of nitrogen can be removed in the algae-bacteria consortium system, whilst only 29% in the algae system and 1% in the bacteria system (Liang et al. 2013). So far, the mechanism of this mutualistic interaction has not been identified. One hypothesis is that bacteria excrete hormones to stimulate algae for nitrate uptake (Hernandez et al. 2009). Another popular one believes that this is caused by the rapid change of culture conditions rather than direct impact from one strain to the other (He et al. 2013). Due to bacterial glucose uptake, algae need to trigger photosynthesis to fix CO_2, causing the synthesis of relevant pigments (*e.g.* chlorophyll). This enhances algal nitrate uptake. Indeed, previous work has declared that algal chlorophyll *a* content in the consortium is 40% more than that in the single system (Liang et al. 2013).

The current study cannot verify the first theory, as machine learning models cannot evolve new mechanisms that are not trained before (consortium data not used for training). However, as the models are trained by the 4 single strain datasets, they can predict the response of cell growth and nutrient uptake of each strain under different conditions well. The close prediction between Case 3 and consortium data therefore favours the second hypothesis. It is also noticed that although a kinetic model is constructed based on physical observations, it can only test hypotheses which are already included in its structure. In other words, the kinetic model cannot be used to identify an unknown mechanism if it does not have any parameter taking into account this mechanism. In fact, as the kinetic model in this study does not contain parameters representing the effect of bacterial hormones on algal nitrate uptake, it cannot be used to verify the first hypothesis either.

4.2.2 Investigation of algae-bacteria interaction under low nutrient concentrations

Same as above, ANNs and GPs exhibit similar results in Cases 1 and 2 (Figs. 7 (a), (c), (e), (g)), with GPs predicting closer results to the data compared to the ANNs in Case 3 (Figs. 7(b), (f), (h)). Once again, GPs results are chosen for analysis. From Figs. 7(a) and 7(c), it can be seen that algae growth is still significant in this process. However, in this system the algae-bacteria competition becomes severe and acts as the primary interaction. Firstly, uptake of glucose and nitrate in the experiment lies in between Cases 1 and 2 (Figs. 7(c), (g)), suggesting neither algae nor bacteria can grow fully. As Case 1 predicts closer cell growth and nitrate uptake to the experiment (Figs. 7(a), (c)), it is concluded that algae growth slightly prevails in the system. Secondly, the constant underestimation of concentrations of phosphate (Fig. 7(f)) and glucose (Fig. 7(h)) and overestimation of biomass concentration (Fig. 7(b)) in Case 3 suggest a strong competition for multiple nutrients (*i.e.* phosphate and glucose). Thirdly, the high uncertainty estimated by GPs in Case 3 also (Figs. 7(d), (f), (h)) implies that algae and bacteria encounter an unexperienced circumstance, probably caused by their intense competition. Finally, the algae-bacteria mutualistic interaction is not observed in this condition, meaning that this consortium is governed by a rather different mechanism.

5. Conclusion

The algae-bacteria consortium wastewater treatment process is one of the most sophisticated biosystems governed by contradictory mechanisms under different conditions. Constructing an accurate model is time/resource-consuming and

challenging, particularly if the datasets are scarce and incomplete. This work therefore presents a heuristic model selection procedure:

1. A kinetic model should be designed firstly. Classic models can deal with three operating factors, beyond which there is no effective structure and parameter estimation can be an issue;

2. A GP could be more effective than an ANN for scarce datasets. Using the hyperparameter selection framework is vital. A GP requires fewer datasets (up to 5) than an ANN (50-200);

3. Linear interpolation is generally accurate enough to fill missing data. If the system changes dramatically, a kinetic model should be constructed to estimate the missing information;

4. Advanced real-time optimal control frameworks *e.g.* economic model predictive control can be used if accuracy of the designed model is limited due to the scarcity of available data.

Acknowledgement

This project has received funding from the EPSRC project (EP/P016650/1). This project has also received funding from the National Natural Science Foundation of China (No. 21776232).

References

Adesanya, Victoria O., Matthew P. Davey, Stuart A. Scott, and Alison G. Smith. 2014. "Kinetic Modelling of Growth and Storage Molecule Production in Microalgae under Mixotrophic and Autotrophic Conditions." *Bioresource Technology* 157

(April): 293–304.

Bankar, Sandip, Vivek Dhumal, Devshri Bhotmange, Sunil Bhagwat, and Rekha Singhal. 2014. “Empirical Predictive Modelling of Poly-Ɛ-Lysine Biosynthesis in Resting Cells of Streptomyces Noursei.” *Food Science and Biotechnology* 23 (1): 201–7.

Bradford, Eric, Artur M. Schweidtmann, Dongda Zhang, Keju Jing, and Ehecatl Antonio del Rio-Chanona. 2018. “Dynamic Modeling and Optimization of Sustainable Algal Production with Uncertainty Using Multivariate Gaussian Processes.” *Computers & Chemical Engineering*, August.

Delgadillo-Mirquez, Liliana, Filipa Lopes, Behnam Taidi, and Dominique Pareau. 2016. “Nitrogen and Phosphate Removal from Wastewater with a Mixed Microalgae and Bacteria Culture.” *Biotechnology Reports* 11 (September): 18–26.

Dineshkumar, R., Gunaseelan Dhanarajan, Sukanta Kumar Dash, and Ramkrishna Sen. 2015. “An Advanced Hybrid Medium Optimization Strategy for the Enhanced Productivity of Lutein in Chlorella Minutissima.” *Algal Research* 7 (January): 24–32.

Hart, William E., Carl Laird, Jean-Paul Watson, and David L. Woodruff. 2012. *Pyomo – Optimization Modeling in Python*. Vol. 67. Springer Optimization and Its Applications. Boston, MA: Springer US.

Harun, Irina, Ehecatl Antonio Del Rio-Chanona, Jonathan L. Wagner, Kyle J. Lauersen, Dongda Zhang, and Klaus Hellgardt. 2018. “Photocatalytic Production of Bisabolene from Green Microalgae Mutant: Process Analysis and Kinetic

Modeling." *Industrial & Engineering Chemistry Research*, July, acs.iecr.8b02509.

He, P.J., B. Mao, F. Lü, L.M. Shao, D.J. Lee, and J.S. Chang. 2013. "The Combined Effect of Bacteria and Chlorella Vulgaris on the Treatment of Municipal Wastewaters." *Bioresource Technology* 146 (October): 562–68.

Hernandez, Juan-Pablo, Luz E. De-Bashan, D. Johana Rodriguez, Yaneth Rodriguez, and Yoav Bashan. 2009. "Growth Promotion of the Freshwater Microalga Chlorella Vulgaris by the Nitrogen-Fixing, Plant Growth-Promoting Bacterium Bacillus Pumilus from Arid Zone Soils." *European Journal of Soil Biology* 45 (1): 88–93.

Jia, Huijun, and Qiuyan Yuan. 2016. "Removal of Nitrogen from Wastewater Using Microalgae and Microalgae–bacteria Consortia." Edited by Arno Rein. *Cogent Environmental Science* 2 (1).

Jiao, Kailin, Jingyu Chang, Xianhai Zeng, I-Son Ng, Zongyuan Xiao, Yong Sun, Xing Tang, and Lu Lin. 2017. "5-Aminolevulinic Acid Promotes Arachidonic Acid Biosynthesis in the Red Microalga Porphyridium Purpureum." *Biotechnology for Biofuels* 10 (1): 168.

Jing, Keju, Yuanwei Tang, Chuanyi Yao, Ehecatl Antonio del Rio-Chanona, Xueping Ling, and Dongda Zhang. 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): 371–81.

John, Joel. 2018. "Microbial Fermentation Technology Market."

Kameswaran, Shivakumar, and Lorenz T. Biegler. 2008. “Convergence Rates for Direct Transcription of Optimal Control Problems Using Collocation at Radau Points.” *Computational Optimization and Applications* 41 (1): 81–126.

Kirk, Paul D. W., and Michael P. H. Stumpf. 2009. “Gaussian Process Regression Bootstrapping: Exploring the Effects of Uncertainty in Time Course Data.” *Bioinformatics* 25 (10): 1300–1306.

Liang, Zhijie, Yan Liu, Fei Ge, Yin Xu, Nengguo Tao, Fang Peng, and Minghung Wong. 2013. “Efficiency Assessment and pH Effect in Removing Nitrogen and Phosphorus by Algae-Bacteria Combined System of Chlorella Vulgaris and Bacillus Licheniformis.” *Chemosphere* 92 (10): 1383–89.

Quinn, Jason, Lenneke de Winter, and Thomas Bradley. 2011. “Microalgae Bulk Growth Model with Application to Industrial Scale Systems.” *Bioresource Technology* 102 (8): 5083–92.

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge: The MIT Press.

Rio-Chanona, Ehecatl Antonio del, Nur rashid Ahmed, Dongda Zhang, Yinghua Lu, and Keju Jing. 2017. “Kinetic Modeling and Process Analysis for Desmodesmus Sp. Lutein Photo-Production.” *AIChE Journal* 63 (7): 2546–54.

Rio-Chanona, Ehecatl Antonio del, Pongsathorn Dechatiwongse, Dongda Zhang, Geoffrey C. Maitland, Klaus Hellgardt, Harvey Arellano-Garcia, and Vassilios S. Vassiliadis. 2015. “Optimal Operation Strategy for Biohydrogen Production.” *Industrial & Engineering Chemistry Research* 54 (24): 6334–43.

Rio-Chanona, Ehecatl Antonio del, Fabio Fiorelli, Dongda Zhang, Nur rashid Ahmed, Keju Jing, and Nilay Shah. 2017. “An Efficient Model Construction Strategy to Simulate Microalgal Lutein Photo-Production Dynamic Process.” *Biotechnology and Bioengineering* 114 (11): 2518–27.

Rio-Chanona, Ehecatl Antonio del, Emmanuel Manirafasha, Dongda Zhang, Qian Yue, and Keju Jing. 2016. “Dynamic Modeling and Optimization of Cyanobacterial C-Phycocyanin Production Process by Artificial Neural Network.” *Algal Research* 13 (January): 7–15.

Subashchandrabose, Suresh R., Balasubramanian Ramakrishnan, Mallavarapu Megharaj, Kadiyala Venkateswarlu, and Ravi Naidu. 2011. “Consortia of Cyanobacteria/microalgae and Bacteria: Biotechnological Potential.” *Biotechnology Advances* 29 (6): 896–907.

Tulsyan, Aditya, Christopher Garvin, and Cenk Ündey. 2018. “Advances in Industrial Biopharmaceutical Batch Process Monitoring: Machine-Learning Methods for Small Data Problems.” *Biotechnology and Bioengineering* 115 (8): 1915–24.

Valdez-Castro, L., I. Baruch, and J. Barrera-Cortés. 2003. “Neural Networks Applied to the Prediction of Fed-Batch Fermentation Kinetics of Bacillus Thuringiensis.” *Bioprocess and Biosystems Engineering* 25 (4): 229–33. https://doi.org/10.1007/s00449-002-0296-7.

Vatcheva, I, H de Jong, O Bernard, and N J I Mars. 2006. “Experiment Selection for the Discrimination of Semi-Quantitative Models of Dynamical Systems.” *Artificial Intelligence* 170 (4–5): 472–506.

Wächter, Andreas, and Lorenz T. Biegler. 2006. “On the Implementation of an

Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming." *Mathematical Programming* 106 (1): 25–57.

Wang, Jufang, Meng Lin, Mengmeng Xu, and Shang-Tian Yang. 2015. "Anaerobic Fermentation for Production of Carboxylic Acids as Bulk Chemicals from Renewable Biomass." In *Advances in Biochemical Engineering/Biotechnology*, 323–61.

Wood, Laura. 2018. "Algae Products Market by Type, Application, Source, Form, and Region - Global Forecast to 2023."

Zhang, D., P. Dechatiwongse, E.a. del Rio-Chanona, G.C. Maitland, K. Hellgardt, and V.S. Vassiliadis. 2015. "Modelling of Light and Temperature Influences on Cyanobacterial Growth and Biohydrogen Production." *Algal Research* 9 (May). Elsevier B.V.: 263–74.

Zhang, Dongda, Pongsathorn Dechatiwongse, Ehecatl Antonio Del-Rio-Chanona, Klaus Hellgardt, Geoffrey C. Maitland, and Vassilios S. Vassiliadis. 2015. "Analysis of the Cyanobacterial Hydrogen Photoproduction Process via Model Identification and Process Simulation." *Chemical Engineering Science* 128 (May): 130–46.

Zhang, Dongda, and Vassilios S. Vassiliadis. 2015. "Chlamydomonas Reinhardtii Metabolic Pathway Analysis for Biohydrogen Production under Non-Steady-State Operation." *Industrial & Engineering Chemistry Research* 54 (43): 10593–605.

Table I: Summary of the current experiments.

Experiments for model construction				
Single strain processes	Biomass	Glucose	TN	TP
Exp. 1: Algae in high nutrients con.	0.24 g/L	500 mg/L	120 mg/L	19 mg/L
Exp. 2: Algae in low nutrients con.	0.24 g/L	100 mg/L	19 mg/L	3 mg/L
Exp. 3: Bacteria in high nutrients con.	0.24 g/L	500 mg/L	120 mg/L	19 mg/L
Exp. 4: Bacteria in low nutrients con.	0.24 g/L	100 mg/L	19 mg/L	3 mg/L
Experiments for algae-bacteria consortium wastewater treatment process investigation				
Exp. 5: Consortium in high nutrients con.	0.48 g/L	500 mg/L	120 mg/L	19 mg/L
Exp. 6: Consortium in low nutrients con.	0.48 g/L	100 mg/L	19 mg/L	3 mg/L

Table II: Values of kinetic model parameters for algal and bacterial wastewater treatment processes with high and low nutrients concentration.

Values of parameters for the bacterial kinetic model

Parameter	High con.	Low con.	Parameter	High con.	Low con.
μ, h^{-1}	0.109	0.0821	μ_d, L/(g·h)	0.0854	0.103
K_N, mg/L	0.00860	0.00873	K_C, mg/L	0.0	0.0
K_P, mg/L	0.0	0.001	Y_{C1} mg/g	217.0	85.5
Y_{N1} mg/g	5.36	4.36	Y_{P1} mg/g	2.74	2.47
Y_{C2} mg/(g·h)	0.839	0.172	Y_{N2} mg/(g·h)	0.0559	0.0132
Y_{P2} mg/(g·h)	0.00833	0.00373			

Values of parameters for the algal kinetic model

Parameter	High con.	Low con.	Parameter	High con.	Low con.
μ, h^{-1}	0.329	0.116	X_{max}, g/L	2.70	2.13
K_N, mg/L	15.2	0.010	K_C, mg/L	10.0	76.8

K_P, mg/L	36.9	0.001	Y_{C1} mg/g	20.4	20.4
Y_{N1} mg/g	8.62	8.63	Y_{P1} mg/g	0.829	0.822
Y_{C2} mg/(g·h)	0.0630	0.0217	Y_{N2} mg/(g·h)	0.138	0.0
Y_{P2} mg/(g·h)	5.01	0.00198			

Table III: Total time consumed for model construction.

Algal models			
Time consumption	Kinetic model	ANN	GP
Time for model structure design	8 weeks	3 days	5 days
Time for parameter estimation	84 seconds	246 seconds	162 seconds
Bacterial models			
Time consumption	Kinetic model	ANN	GP
Time for model structure design	5 weeks	2 days	3 days

Time for parameter estimation	16 seconds	262 seconds	101 seconds

Figures

Figure 1: Schematic of ANN and GP. (a): A classic ANN structure. (b): Prior and Posterior distributions of a GP regression. The dashed lines covered region is the prior distribution (initial guess), and the solid lines covered region is the posterior distribution (updated distribution).

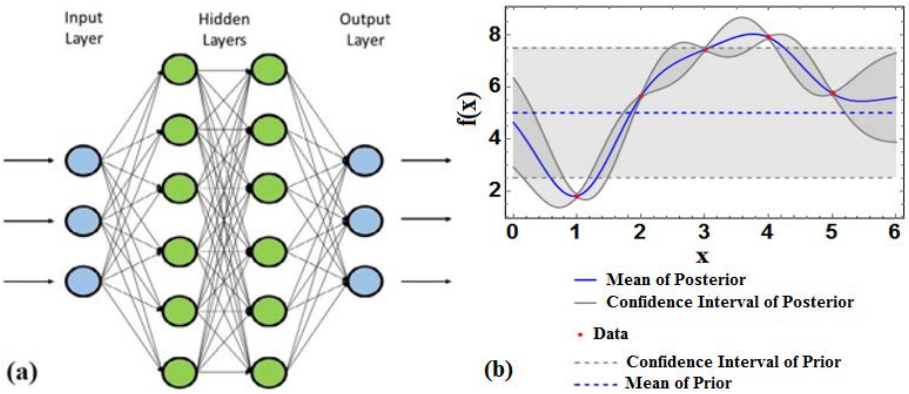


Figure 2: Simulation results of algal wastewater treatment process with high nutrients concentration. (a): Biomass concentration; (b): Nitrate concentration; (c): Glucose concentration; (d): Phosphate concentration. Red point (open circle with cross): experimental data. Open diamond: ANN simulation result. Blue point (filled circle): GP simulation result (the uncertainty is not detectable). Black line: kinetic model simulation result.

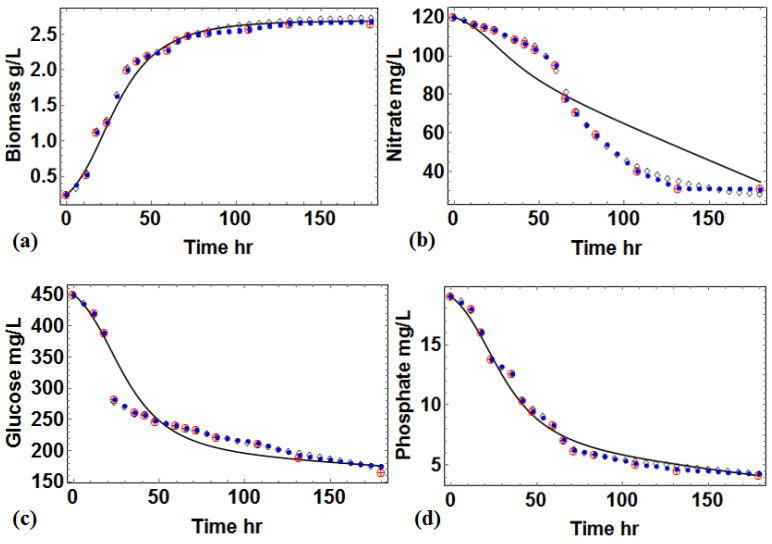


Figure 3: Simulation results of algal wastewater treatment process with low nutrients concentration. (a): Biomass concentration; (b): Nitrate concentration; (c): Glucose concentration; (d): Phosphate concentration. Red point (open circle with cross): experimental data. Open diamond: ANN simulation result. Blue point (filled circle): GP simulation result (the uncertainty is not detectable). Black line: kinetic model simulation result.

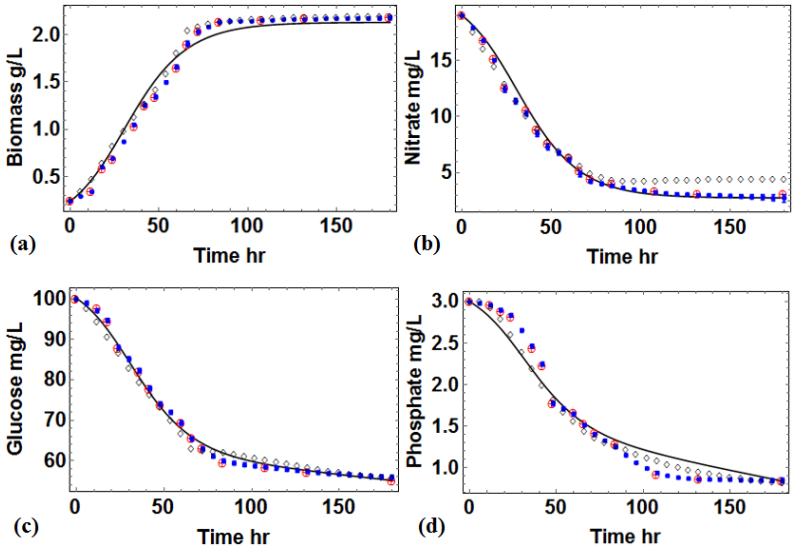


Figure 4: Simulation results of bacterial wastewater treatment process with high nutrients concentration. (a): Biomass concentration; (b): Nitrate concentration; (c): Glucose concentration; (d): Phosphate concentration. Red point (open circle with cross): experimental data. Open diamond: ANN simulation result. Blue point (filled circle): GP simulation result (the uncertainty is not detectable). Black line: kinetic model simulation result.

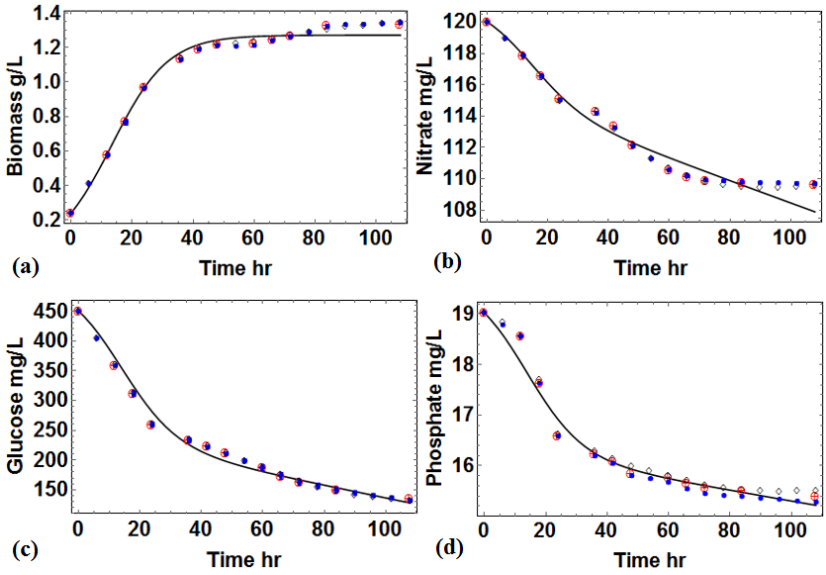


Figure 5: Simulation results of bacterial wastewater treatment process with low nutrients concentration. (a): Biomass concentration; (b): Nitrate concentration; (c): Glucose concentration; (d): Phosphate concentration. Red point (open circle with cross): experimental data. Open diamond: ANN simulation result. Blue point (filled circle): GP simulation result (the uncertainty is not detectable). Black line: kinetic model simulation result.

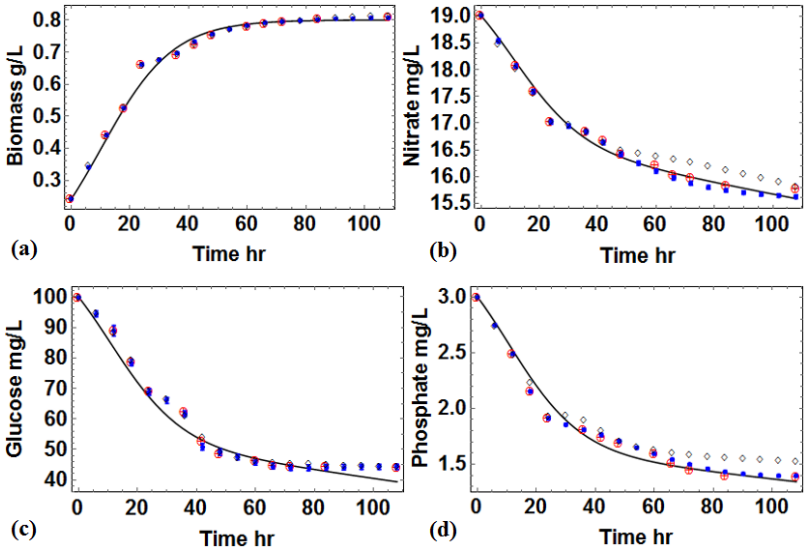


Figure 6: Prediction results of algae-bacteria consortium wastewater treatment process with high nutrients concentration. (a), (c), (e), (g): Prediction results of biomass concentration and nutrients concentration in Case 1 and Case 2. Blue points (open circle with cross): Experimental data. Filled circles: GP prediction results of Case 1 (red circle) and Case 2 (black circle). Open circles: ANN prediction results of Case 1 (red circle) and Case 2 (black circle). (b), (d), (f), (h): Prediction results of biomass concentration and nutrients concentration in Case 3. Blue points (open circle with cross): Experiment data. Filled circles: GP prediction result. Open circles: ANN prediction result.

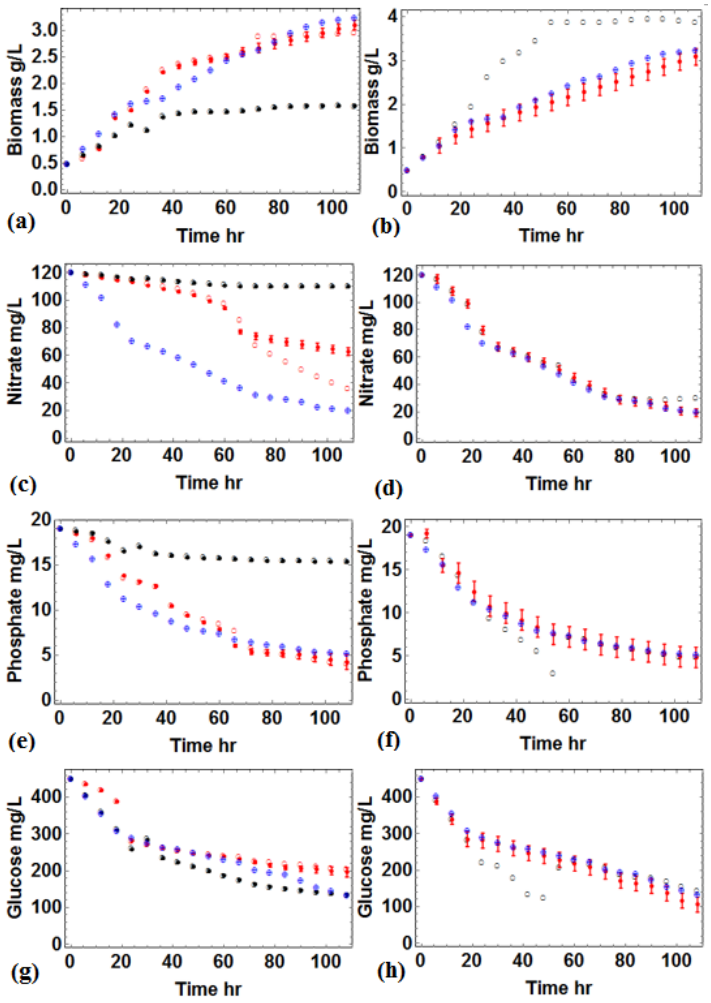


Figure 7: Prediction results of algae-bacteria consortium wastewater treatment process with low nutrients concentration. (a), (c), (e), (g): Prediction results of biomass concentration and nutrients concentration in Case 1 and Case 2. Blue points (open circle with cross): Experimental data. Filled circles: GP prediction results of Case 1 (red circle) and Case 2 (black circle). Open circles: ANN prediction results of Case 1 (red circle) and Case 2 (black circle). (b), (d), (f), (h): Prediction results of biomass concentration and nutrients concentration in Case 3. Blue points (open circle with cross): Experiment data. Filled circles: GP prediction result. Open circles: ANN prediction result.