An efficient model structure identification strategy for bioprocess hybrid modelling

Dongda Zhang[1], Thomas Savage[1], Bovinille Anye Cho[1], and Ehecatl Del Rio-Chanona[2]

[1]University of Manchester
[2]Imperial College London

May 5, 2020

Abstract

Integrating physical knowledge with machine learning is critical to developing industrially-focused digital twins for monitoring and optimisation of biomanufacturing systems. However, identifying the correct model structure to quantify kinetic mechanisms poses a challenge for the construction of mechanistic and data-driven models. This study proposes a hybrid modelling strategy comprising of a simple kinetic model to describe the overall process trajectory and a data-driven model to estimate the mismatch between the kinetic equations and real process. An automatic model structure identification algorithm is used to identify the most probable kinetic model structure and minimum number of data-driven model parameters that can well represent different bioprocess behaviours over broad operating conditions. Through this approach, a hybrid model was constructed to simulate biomass growth, nutrient consumption, and product synthesis in an algal photo-production process. Performance of this model for predictive modelling, optimisation, and online self-calibration is demonstrated, indicating its advantages for industrial application.

1. Introduction

Developing industrially focused mathematical models is one of the grand research challenges for the design, operation and commercialisation of next generation sustainable chemical and biochemical processes. Due to the lack of petroleum resources and the severe environmental issues surrounding them, microorganism based bio-production processes have become an attractive candidate to substitute traditional chemical processes for the industrial synthesis of platform chemicals and high-value materials [1–3]. Given the sophisticated metabolisms, two characteristics exist in most bio-production processes. The first is that different strains and species share similar behaviour with respect to biomass growth, nutrient consumption, and bioproduct accumulation due to their delicate metabolic regulation mechanisms[4,5]. Whilst the second is that bioprocesses are difficult to reproduce, meaning that their performance varies from batch to batch even under similar operating conditions, as metabolic reactions are sensitive to the change of culture environment[6,7].

At this moment, different predictive models have been proposed to account for bioprocess complexities. On the one hand, elaborate kinetic models have been developed by embedding new physical understandings into classic models such as the Monod and the Droop model[8,9]. These have been used to simulate, optimise, and scale up both fermentation processes and algal photo-production systems[7,10]. However, identifying a correct model structure to quantify the physical knowledge is a challenging task, usually with long development times. This often results in a complex model structure leading to issues with parameter estimation and identifiability, and sacrificing the model's predictive capability [11]. On the other hand, frontier machine learning models such as artificial neural networks, Gaussian processes, and reinforcement learning have been applied for bioprocess dynamic modelling and online optimisation, and their competency has been reported in a number of publications[12–14]. Although these data-driven models can well capture complex process behaviours in a specific operating range without prior physical knowledge, they suffer from other inherent weaknesses, such as the risk in model overfitting and difficulties in extrapolating a broader range of metabolism governed process behaviours[11,15].

To resolve these challenges, a third modelling strategy – hybrid modelling – has been proposed in recent years [16]. This strategy aims to combine physical knowledge and machine learning into a hybrid model structure to inherit the respective advantages of both kinetic models and data-driven models. The structure of a hybrid model is flexible (*e.g.* parallel structure or sequential structure) and depends on the amount of available physical information and process data [17]. In spite of its merits and industrial potential, there exists only a few pioneering research studies attempting to improve and apply this technology into bioprocess engineering [18–20]. In addition, hybrid model identification remains a challenge, as its kinetic aspect suffers from difficulties in quantifying physical knowledge and its data-driven part poses risks in overfitting. As a result, this study aims to develop a general framework that integrates state-of-the-art automatic model structure identification technology into the hybrid modelling strategy to facilitate its future industrial applications in bioprocess engineering.

2. Methodology

2.1 Computational experiment setup

In order to illustrate the proposed model construction framework and to test the efficiency of the hybrid model in process predictive modelling and optimisation, computational experiments were used in this work so that the best process performance can be identified as the benchmark. A microalgal lutein production process was selected as the specific case study, as an algal photo-production process is biologically more complex than a conventional fermentation process. Algal biomass growth and lutein synthesis are mainly affected by light intensity and nitrate concentration [21]. A complex kinetic model designed in our previous work was used to generate computational experimental data for different purposes and is presented in Eq. (1a)-(1d)[22]. This model can well simulate effects of light intensity, light attenuation, and nitrate supply on biomass growth and lutein production. However, given its complex model structure, its application in process optimisation and bioreactor design is limited, and identifying its model structure is time consuming.

$$\frac{dc_X}{\mathrm{dt}} = \frac{u_m}{20} \bullet \left(\frac{I_0}{I_0 + k_s + \frac{I_0^2}{k_i}} + \sum_{n=1}^{9} \frac{2 \bullet I_n}{I_n + k_s + \frac{I_n^2}{k_i}} + \frac{I_{10}}{I_{10} + k_s + \frac{I_{10}^2}{k_i}} \right) \bullet \frac{c_N}{c_N + K_N} \bullet c_X - u_d \bullet c_X$$

$$\frac{dc_N}{\mathrm{dt}} = -Y_{N/X} \bullet \frac{u_m}{20} \bullet \left(\frac{I_0}{I_0 + k_s + \frac{I_0^2}{k_i}} + \sum_{n=1}^{9} \frac{2 \bullet I_n}{I_n + k_s + \frac{I_n^2}{k_i}} + \frac{I_{10}}{I_{10} + k_s + \frac{I_{10}^2}{k_i}} \right) \bullet \frac{c_N}{c_N + K_N} \bullet c_X + F_{\mathrm{in}} \bullet c_{N,in}$$

$$\frac{dc_L}{\mathrm{dt}} = \frac{k_m}{20} \bullet \left(\frac{I_0}{I_0 + k_{\mathrm{sL}} + \frac{I_0^2}{k_{\mathrm{iL}}}} + \sum_{n=1}^{9} \frac{2 \bullet I_n}{I_n + k_{\mathrm{sL}} + \frac{I_n^2}{k_{\mathrm{iL}}}} + \frac{I_{10}}{I_{10} + k_{\mathrm{sL}} + \frac{I_{10}^2}{k_{\mathrm{iL}}}} \right) \bullet \frac{c_N}{c_N + K_{\mathrm{NL}}} \bullet c_X - k_d \bullet c_L \bullet c_X$$

$$I(l) = I_0 \bullet \left(e^{-(\tau \bullet c_X + K_a) \bullet l} + e^{-(\tau \bullet c_X + K_a) \bullet (z-l)} \right) \tag{1d}$$

where c_X, c_N, and c_L are the concentrations of biomass, nitrate, and lutein, respectively. F_{in} and$c_{N,in}$ are nitrate inflow rate and concentration, respectively.l is the distance from light source, z is width of the reactor (0.084 m). I_n is the local light intensity at a distance of$\frac{n \bullet z}{10}$ $(n = 1, \ldots, 10)$ m away from the incident light surface area, while I_0 is the incident light intensity. Parameter values and their physical meanings can be found in[22].

This complex model served to act as the "true" experimental system. Initially, three computational experiments were conducted under a broad spectrum of operating conditions from nitrate-limiting conditions and photo-limitation to nitrate-excessive conditions and photo-inhibition. Data generated in these experiments was then used to construct the hybrid model. Once constructed, the hybrid model was exploited to predict and optimise a number of fed-batch processes under different conditions, and computational experimental verification using this complex model was executed to verify accuracy of the hybrid model. Detailed operating conditions of the computational experiments are listed in Table 1.

Table 1: Operating conditions of all the computational experiments: Exp. 1-3 (batch processes), used for parameter estimation; Offline (fed-batch processes): initial conditions of the four offline optimisation processes; Online (fed-batch process): initial conditions of the model self-calibration experiment.

	Exp. 1	Exp. 2	Exp. 3	Offline	Online
Initial biomass concentration (g L^{-1})	0.2	0.2	0.2	0.2	0.2
Incident light intensity (µmol m^{-2} s^{-1})	100	450	800	100	100
Initial nitrate concentration (mg L^{-1})	600	200	1000	600	100
Initial lutein production (mg L^{-1})	0.0	0.0	0.0	0.0	0.0
Operating time (h)	120	120	120	120	120

2.2 Hybrid model construction

As stated previously, the complexity and nonlinearity of a bioprocess consists both of identified physical mechanisms and undetermined process dynamics. Therefore, as shown in Eq. 2, the principle of a hybrid model is to quantify the underlying bioprocess behaviour by using a kinetic model (first term on the right-hand side of Eq. (2a)) to tackle the known dynamics (physical knowledge) and a data-driven model (second term on the right-hand side of Eq. (2a)) to account for the unknown dynamics.

$$\frac{d\mathbf{S}}{\mathrm{dt}} = K(\mathbf{S}) + D(\mathbf{S}) \tag{2a}$$

$$D(S_i) = a_i \bullet S_i + \sum_{j=1}^{M} a_{\mathrm{ij}} \bullet S_i \bullet S_j \tag{2b}$$

where $\mathbf{S}$ is the state variable vector$\mathbf{S} = (X, N, P)^T$,and X, N, P represent the concentrations of biomass, nutrient, and product respectively.$K(\mathbf{S})$ and $D(\mathbf{S})$ are the kinetic model and the data-driven model, respectively. S_i is a state variable, M is the total number of state and control variables, a_i and a_{ij} are coefficients of the polynomial terms.

Distinct from a pure kinetic model, the kinetic model used for hybrid model construction does not require a complex model structure to fully capture the process nonlinearity; it only aims to approximate the overall trend of process dynamics. Thus, classic kinetic models such as the Monod model and the Droop model can be directly adopted without the necessity of further modification based on more detailed physical information. Similarly, compared to a pure data-driven model, the data-driven model used in a hybrid model simulates only the unknown terms, in other words, mismatch between the kinetic model and the process. The nonlinearity of this mismatch is greatly reduced compared to the original bioprocess, as the general behaviour has been described by the kinetic model. A number of previous studies have shown that after subtracting the mechanistic process trajectory, even a simple data-driven model *e.g.* PLS which is mainly used for linear systems, can well capture the processes physically undetermined behaviour [11]. As a result, it may not be necessary to embed sophisticated data-driven models such as ANNs or GPs into a hybrid model.

In this study, a 2nd degree polynomial regression model shown as Eq. (2b) was selected as the data-driven model to estimate the mismatch between the kinetic model and the process data. 2nd degree polynomial regression models have been predominantly used in response surface methodology for optimal experimental design and analysis [23,24]. Their use has been extended into dynamic systems through the recent progress in dynamic response surface methodology [25]. Moreover, a 2nd degree dynamic polynomial model is also known as an extension of the Lotka-Volterra model [26], a classic model used in bioinformatics to simulate growth and competition amongst different populations. Parameter estimation and uncertainty analysis of a polynomial regression model is more straightforward to implement when compared to a complex data-driven model, as extracting gradient information is challenging in ANNs or GPs, as is their parameter estimation and optimisation. This feature is particularly advantageous for industrial applications, as estimating bioprocess uncertainty is a severe challenge for industrial systems operation and decision-making.

When considering the kinetic part $K(\mathbf{S})$, its aim is to approximate the process trajectory. As the current system is mainly affected by light intensity, light attenuation, and nitrate supply, only these three factors are considered when building the simple kinetic model. Photo-inhibition, biomass decay, and lutein self-degradation are not included in $\mathrm{K}(\mathbf{S})$ as their effects are subtle and will result in a highly complex model structure such as Eq. (1a)-(1f). They are characterised as the model-plant mismatch and are considered by the data-driven model. The kinetic part of the hybrid model is presented in Eq. (3a)-(3d). It is worth noticing that specific to algal photo-production systems, although light intensity affects cell growth which in turn influences nitrate uptake, it has yet been confirmed if light directly triggers nitrate consumption. In fact, several previously proposed models do not link light intensity with nitrate uptake (denoted by Eq. (3b)) [27,28], whilst others tried to establish a direct interaction between these two (seen as Eq. (3c)) [29,30]. Similarly, it is still not clear how significantly light affects lutein synthesis. Thus, Eq. (3d) and Eq. (3e) are proposed in this study based on different hypothesises (no/weak effect and strong effect, respectively). The final kinetic model structure will be determined via the automatic model identification framework as presented in the next section.

$$\left.\frac{dc_X}{\mathrm{dt}}\right|_K = u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X \tag{3a}$$

$$\left.\frac{dc_N}{\mathrm{dt}}\right|_K = -Y_{N/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet c_X + F_{\mathrm{in}} \bullet c_{N,in} \tag{3b}$$

$$\left.\frac{dc_N}{\mathrm{dt}}\right|_K = -Y_{N/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X + F_{\mathrm{in}} \bullet c_{N,in} \tag{3c}$$

$$\left.\frac{dc_L}{\mathrm{dt}}\right|_K = Y_{L/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet c_X \tag{3d}$$

$$\left.\frac{dc_L}{\mathrm{dt}}\right|_K = Y_{L/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_{\mathrm{sL}}} \bullet c_X \tag{3e}$$

where the subscript K refers to the kinetic part of the hybrid model.

2.3 Automatic model structure identification

Two challenges are encountered when constructing a hybrid model, the first being identifying the physically correct model structure of the kinetic part, and the second being preventing the overfitting of the data-driven part. For a purely kinetic model, the first challenge is addressed via model discrimination which consists of several iterations to develop various structures and test their data fitting performance[31]. For a purely data-driven model, the second challenge is solved via a hyper-parameter selection framework to identify the simplest model structure that well fits the data according to certain criteria [32]. In this work, these two challenges are addressed simultaneously for the first time by creating an automatic model structure detection strategy comprised of model reformulation and sparse optimisation. The original idea of this strategy was proposed to reveal natural laws based on experiment data using basic mathematical operators without any prior knowledge[33]. This idea was then applied to nonlinear dynamic systems in 2016 to identify the governing equations for fluid dynamics[34]. The current study launched the first investigation to modify this strategy and adopt it into bioprocess hybrid modelling. Details of this strategy are explained below.

Initially, the original hybrid model is reformulated into a mixed-integer nonlinear programming (MINLP) problem by assigning binary variables into each individual term in the hybrid model, as presented in Eq. (4a)-(4c). By adding two equality constraints Eq. (4d)-(4e), different kinetic equations for nitrate uptake and lutein production are merged into one expression shown as Eq. (4b) and (4c), respectively. As binary variables can only be either 0 or 1, imposing the equality constraints will ensure that only one of the possible kinetic formulations (*e.g.* the first term on the right-hand side of Eq. (3b) or (3c)) is active, with the others being inactive as their binary variables are 0. Through this way, kinetic model discrimination can be executed more efficiently than the trial and error approach. This method can also be applied to general cases in which multiple contradictory kinetic hypotheses are present. Data nominalisation must be executed before using the polynomial regression model. However, this is not needed for kinetic model construction, as each parameter in the model has its unique physical meaning and unit.

$$\frac{dc_X}{\mathrm{dt}} = u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X + b_{10} \bullet a_{10} \bullet c_1 + \sum_{j=1}^{4} b_{1j} \bullet a_{1j} \bullet c_1 \bullet c_j \qquad (4a)$$

$$\frac{dc_N}{\mathrm{dt}} = b_{N1} \bullet \left(-Y_{N/X} \bullet \frac{u_0 \bullet c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X \right) + b_{N2} \bullet \left(-Y_{N/X} \bullet \frac{u_0 \bullet c_N}{c_N + K_N} \bullet c_X \right) + F_{\mathrm{in}} \bullet c_{N,in} + b_{20} \bullet a_{20} \bullet c_2 + \sum_{j=1}^{4}$$

$$\frac{dc_L}{\mathrm{dt}} = b_{L1} \bullet \left(Y_{L/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_{\mathrm{sL}}} \bullet c_X \right) + b_{L2} \bullet \left(Y_{L/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet c_X \right) + b_{30} \bullet a_{30} \bullet c_3 + \sum_{j=1}^{4} b_{3j} \bullet$$

$$b_{N1} + b_{N2} = 1 \qquad (4d)$$

$$b_{L1} + b_{L2} = 1 \qquad (4e)$$

where b_{Ni} and b_{Li} $(i = 1, 2)$ are binary variables for the kinetic model, b_{ij} $(i = 1, 2, 3, j = 0, 1, 2, 3)$ are binary variables for the data-driven model,a_{ij} $(i = 1, 2, 3,\ j = 0, 1, 2, 3, 4)$ are coefficients of the polynomial terms, c_i $(i = 1, 2, 3, 4)$ are normalised values of concentrations of biomass, nitrate, and lutein and incident light intensity, respectively.

Once reformulated, the next step is to implement sparse optimisation to automatically identify the most probable kinetic model structure and minimum number of polynomial terms that can well fit the multiple generated datasets. Therefore, the objective function for parameter estimation is designed such that in addition to minimising residues between model prediction and experimental data, it also penalises the total number of active binary variables in the polynomial terms to avoid overfitting. In this way, the optimal hybrid model structure alongside its parameter values can be simultaneously identified. The objective function therefore is formulated as Eq. 5.

$$\min F = \sum_{k=1}^{n}\sum_{p=1}^{m}\sum_{i=1}^{3}\left(c_{S_i,p,k,E} - c_{S_i,p,k,M}\right)^2 \bullet w_{S_i,j,k} + w_b \bullet \sum_{j=0}^{4}\sum_{i=1}^{3} b_{\mathrm{ij}} \tag{5}$$

where $c_{S_i,p,k,E}$ and $c_{S_i,p,k,E}$ are experimental and model simulated value for concentration of state S_i at time step$p \in m$ in the k^{th} dataset (total number of datasets is n), respectively, $w_{S_i,j,k}$ and w_b are the weight for each data point and the sum of binary variables, respectively.

In this work, parameter estimation was formulated as a weighted nonlinear least squares optimisation problem. However, as the current model is highly nonlinear due to its complex model structure and many discrete (binary) variables, it must be relaxed so that it can be solved using the standard dynamic parameter estimation method[22]. Conventionally, an MINLP problem is relaxed via introducing new parameters and extra linear constraints to reduce nonlinearity of the original problem [35]. For example, converting Eq. (4a) to Eq. (6a) where new continuous parametersB_{1j} are introduced to substitute the bilinear term$b_{1j} \bullet a_{1j}$ and satisfy the constraints Eq. (6b)-(6c).

$$\frac{dc_X}{\mathrm{dt}} = u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau\bullet c_X\bullet z}}{I_0 e^{-\tau\bullet c_X\bullet z} + k_s} \bullet c_X + B_{10} \bullet c_1 + \sum_{j=1}^{4} B_{1j} \bullet c_1 \bullet c_j \tag{6a}$$

$$a_{1j} - a_{1j}^{U} \bullet (1 - b_{1j}) \leq B_{1j} \leq a_{1j} - a_{1j}^{L} \bullet (1 - b_{1j}) \tag{6b}$$

$$a_{1j}^{L} \bullet b_{1j} \leq B_{1j} \leq a_{1j}^{U} \bullet b_{1j} \tag{6c}$$

where a_{1j}^{L} and a_{1j}^{U} are the lower and upper bound ofa_{1j}, respectively.

In spite of its uses over a range of chemical engineering research, this relaxation still results in an MINLP. Given this hybrid dynamic model is highly nonlinear, another relaxation is proposed in this study to reformulate the MINLP into an NLP (nonlinear programming problem) by converting binary variables into continuous variables with two extra constraints Eq. (7a)-(7b). To satisfy constraints Eq. (7a)-(7b), values of b_{ij} have to be either 0 or 1. In this way, the original problem can be effectively solved via the continuous dynamic parameter estimation method [22].

$$0 \leq b_{\mathrm{ij}} \leq 1 \tag{7a}$$

$$b_{\mathrm{ij}} \bullet (1 - b_{\mathrm{ij}}) \leq 0 \tag{7b}$$

The NLP model was discretised by orthogonal collocation over finite elements in time, and was then solved using the interior point nonlinear optimisation solver IPOPT through a multi-start framework. This was executed in the Python optimisation environment Pyomo[36]. The model was simulated in Mathematica 11. Details regarding the this dynamic model parameter estimation procedure can be found in [22].

3. Result and Discussion

3.1 Result of hybrid model construction

Through automatic model structure identification, the hybrid model is identified as Eq. (8a)-(8c) (all inactive terms are removed; active binary variables are not shown as they are equal to 1). Table 2 shows the parameter estimation result. Fig. 1 shows the modelling fitting result.

$$\frac{dc_X}{\mathrm{dt}} = u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X + a_{10} \bullet c_1 + a_{14} \bullet c_1 \bullet c_4 \tag{8a}$$

$$\frac{dc_N}{\mathrm{dt}} = -Y_{N/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_s} \bullet c_X + F_{\mathrm{in}} \bullet c_{N,in} + a_{20} \bullet c_2 \tag{8b}$$

$$\frac{dc_L}{\mathrm{dt}} = Y_{L/X} \bullet u_0 \bullet \frac{c_N}{c_N + K_N} \bullet \frac{I_0 e^{-\tau \bullet c_X \bullet z}}{I_0 e^{-\tau \bullet c_X \bullet z} + k_{\mathrm{sL}}} \bullet c_X + a_{31} \bullet c_3 \bullet c_1 + a_{33} \bullet c_3 \bullet c_3 \tag{8c}$$

From the figure, it is seen that the hybrid model can accurately fit multiple datasets collected over a broad spectrum of operating conditions (*e.g.* substrate limiting to substrate excessive, low light intensity to high light intensity). In addition, the total number of binary variables assigned to the data-driven model is 15, whilst only 5 are active after solving the parameter estimation problem. The correct kinetic expression is also identified more efficiently than traditional model discrimination. Compared to the complex kinetic model shown in Eq. (1a)-(1d), the current hybrid model well fits all the datasets with a simpler model structure.

Table 2: Parameter values of the current hybrid model

Parameter	Value	Parameter	Value
u_0, h^{-1}	0.045	$Y_{N/X}$, mg g^{-1}	315.2
K_N, mg L^{-1}	6.30	$Y_{L/X}$, mg g^{-1}	4.44
K_s, μmol m^{-2} s^{-1}	16.7	K_l, μmol m^{-2} s^{-1}	155.6
τ, m^2 g^{-1}	0.031	a_{10}	-0.0067
a_{14}	0.0024	a_{20}	0.216
a_{31}	0.012	a_{33}	-0.0055

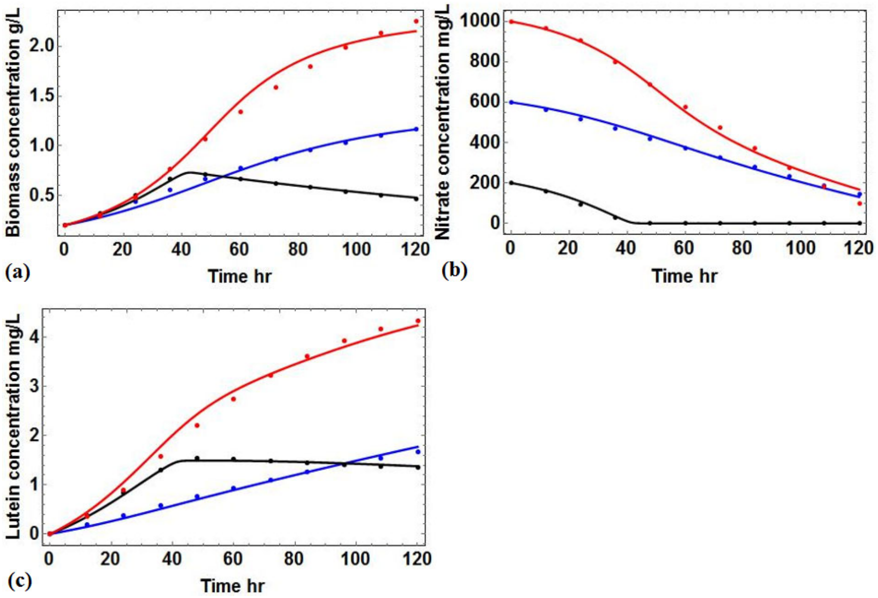


Figure 1: Curve fitting result of the hybrid model. Lines are the hybrid model fitting results, while points are computational experimental data. Blue, black, and red colours refer to Experiments 1, 2, and 3, respectively. (a): biomass concentration; (b): nitrate consumption; (c): lutein production.

3.2 Open-loop optimal control (offline optimisation)

To further assess advantages of this hybrid model, it is first used in offline optimisation to test its ability in long-term process predictive modelling and optimisation. The objective function is to maximise the final lutein production over a 5-day fed-batch operation system in which incident light intensity (100 µmol m^{-2}s^{-1} to 800 µmol m^{-2}s^{-1}) and nitrate inflow rate (0 to 17.5 mL h^{-1} with a concentration of 0.1 M so that total added liquid volume is negligible) can change once per day. Once optimal control actions were estimated by the hybrid model, they were implemented and verified in the computational experiment (using the original complex kinetic model). In addition, the original model was also used to search for the theoretical maximum lutein production in this fed-batch process. This is used as a benchmark to evaluate the hybrid model's predictive capability. Initial operating conditions of this process are listed in Table 1.

Fig. 2 and Fig. 3 shows the model prediction result and experimental verification results. From Fig. 2, it can be observed that the hybrid model can well predict the dynamic process behaviour over the entire operating time course before executing the process. This directly speaks of the predictive capability of the hybrid model. Furthermore, through (computational) experimental verification, the hybrid model is found to provide the same lutein production as the maximum theoretical value estimated by the original kinetic model, indicating its high performance in offline optimisation. The two processes (designed based on the hybrid model: Fig. 2(a), 2(e), 3(c); designed based on the original model: Fig. 2(b), 2(f), 3(d)) have the same trajectory for biomass growth, lutein production, and control actions of incident light intensity. However, culture nitrate concentration (Fig. 2(b) vs. Fig. 2(d)) and nitrate inflow rate (Fig. 3(a) vs. Fig. 3(b)) deviate substantially between the two processes. This indicates that this process is not sensitive to the change of nitrate concentration as long as it is sufficient (400 to 900 mg L^{-1}). It also suggests that multiple optimal solutions exist for this offline optimisation problem.

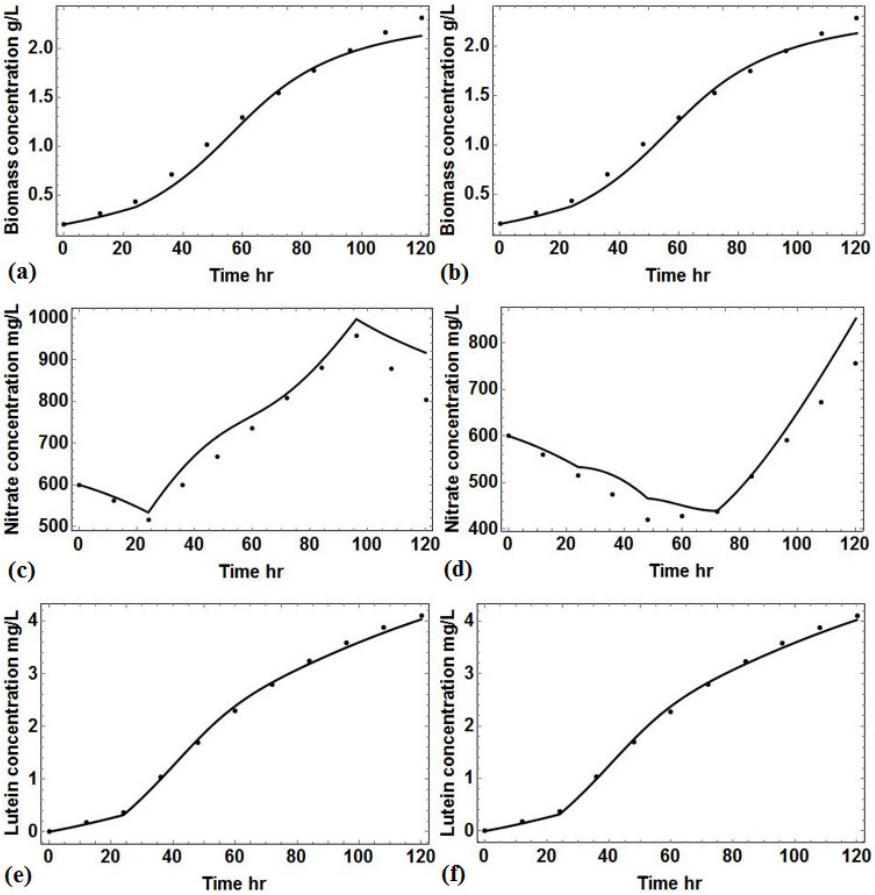


Figure 2: Offline optimisation of the two fed-batch operation processes without practical constraints. Lines are hybrid model prediction results, points are computational experimental verification. (a), (c), and (e): control actions of this fed-batch process as predicted by the hybrid model. (b), (d), and (f): control actions of this fed-batch process as predicted by the original kinetic model (theoretical maximum lutein production), and the hybrid model is used to predict the process behaviour.

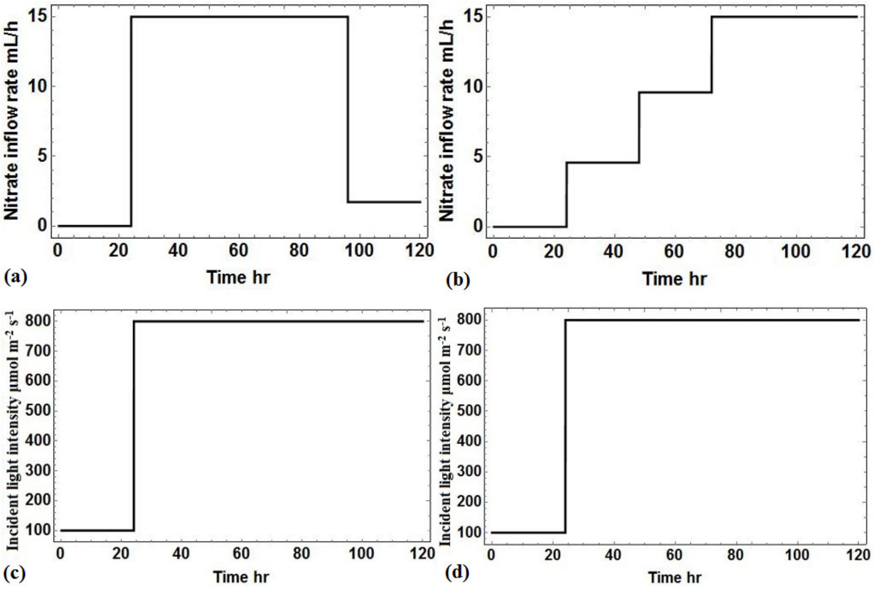


Figure 3: Optimal control actions of the two fed-batch operation processes without practical constraints. (a) and (c): control actions of this process are predicted by the hybrid model. (b) and (d): control actions of this process are predicted by the original kinetic model.

To restrict the solution space and include more practical concerns, a new offline optimisation problem is formulated by embedding three constraints: the first being that changes between two adjacent nitrate inflow rates must be less than 2 mL h^{-1}; the second being that final culture nitrate concentration must be lower than 600 mg L^{-1}; and the third being that total lutein production must be no lower than 4.0 mg L^{-1}. The first constraint aims to prevent drastic changes of culture environment, the second aims to reduce raw material cost, and the third aims to guarantee a similar final production to the theoretical maximum value. The initial operating conditions are set to be the same as before. Through (computational) experimental verification as shown in Fig. 4 and Fig. 5, it is concluded that, once again, the hybrid model well predicted the dynamics of biomass growth and lutein production. Nonetheless, its prediction regarding nitrate concentration (Fig. 4(d)) is found to have large deviations when simulating the experiment designed by the original kinetic model. This means that directly using hybrid model for long-term process prediction, monitoring and optimisation may still be unreliable without adequate online self-calibration mechanism.

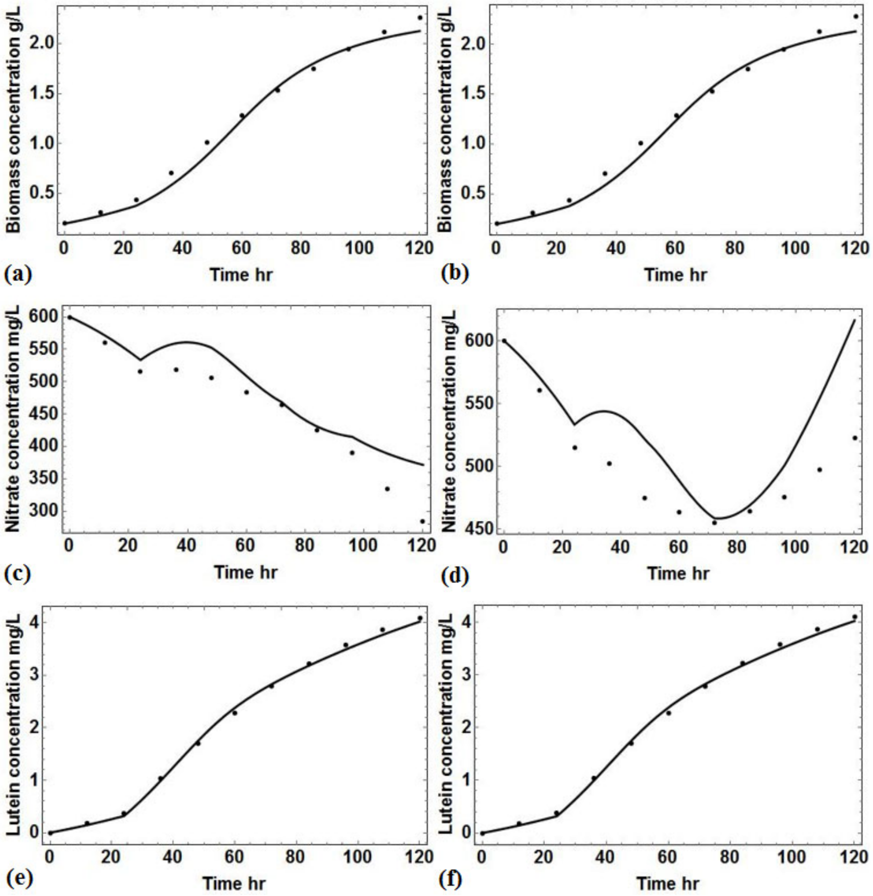


Figure 4: Offline optimisation of the two fed-batch operation processes with the three extra practical constraints. Lines are hybrid model prediction results, points are computational experimental verification. (a), (c), and (e): control actions of this fed-batch process are predicted by the hybrid model. (b), (d), and (f): control actions of this fed-batch process are predicted by the original kinetic model (theoretical maximum lutein production), and the hybrid model is used to predict the process behaviour.

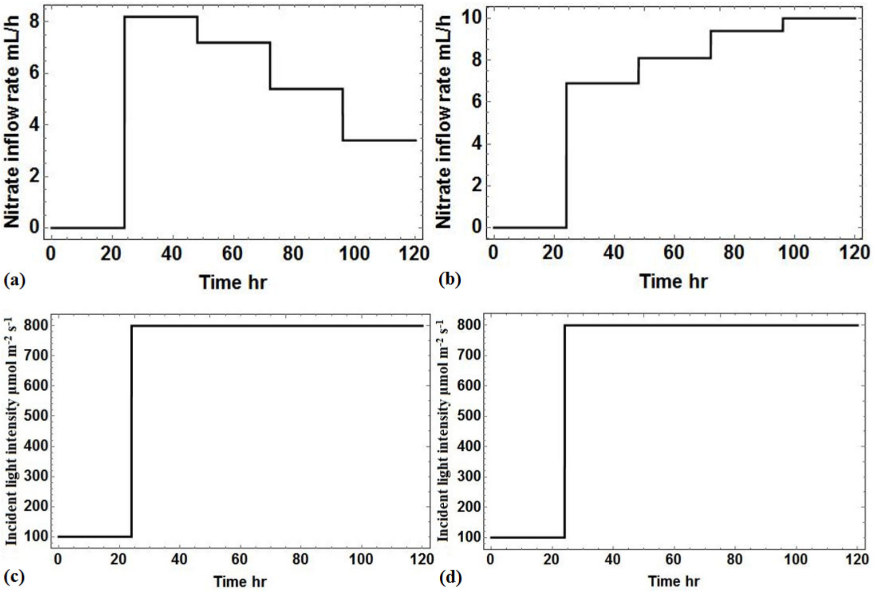


Figure 5: Optimal control actions of the two fed-batch operation processes with the three extra practical constraints. (a) and (c): control actions of this process are predicted by the hybrid model. (b) and (d): control actions of this process are predicted by the original kinetic model.

3.3 Online calibration for state estimation

In practice, new measurement of state variables can be collected through regular sampling and offline analysis during an ongoing process. However, it is not easy to directly use this data to effectively calibrate a kinetic model or a machine learning model. For instance, although parameters of a kinetic model can be rapidly re-estimated, its prediction will always suffer from model-plant mismatch due to the fixed model structure. Although a data-driven model may eliminate model-plant mismatch, it requires substantial amount of data to frequently update its large number of parameters. However, new data collected from an ongoing process is usually limited, hence not able to satisfy this prerequisite. Nonetheless, both of these issues can be resolved by a hybrid model. The kinetic part of a hybrid model does not have to be frequently calibrated as its mismatch has been included in the data-driven part. The data-driven part of a hybrid model has a simple structure with a minimal amount of parameters (*e.g.* 5 parameters in this case), thus new data (usually sampled once per 4 hours in a large scale process, hence 6 data points per day) collected from an ongoing process can be used to effectively update the model's accuracy. As a result, online calibrating a hybrid model is straightforward.

To illustrate this advantage, another case study was conducted. Since the current hybrid model is accurate for long-term offline optimisation, it was decided to first add errors in its parameter values so that a large initial model-plant mismatch is observed. Therefore, a 15% random error was assigned to all the parameters, and the new model's prediction is shown in Fig. 6 when only the initial operating conditions and pre-determined control actions are given. It is seen that large mismatch exists at a later stage of the process. However, after model re-calibration at the end of day 1 (using 6 new data points), it is found that the model's prediction is improved significantly, particularly for biomass and lutein concentrations. Daily calibration of the data-driven part of the model leads to moderate improvement until day 3, beyond which this effect becomes negligible. Nonetheless, it is also observed that the hybrid model still contains an error for online state estimation and long-term prediction, particularly for nitrate concentration. This is because the kinetic part of the model comprises a 15% error in its parameters and has never been calibrated. As a result, once enough new data points are accumulated from the ongoing process, they should be used to re-estimate values of all the parameters in the hybrid model to further improve model accuracy.

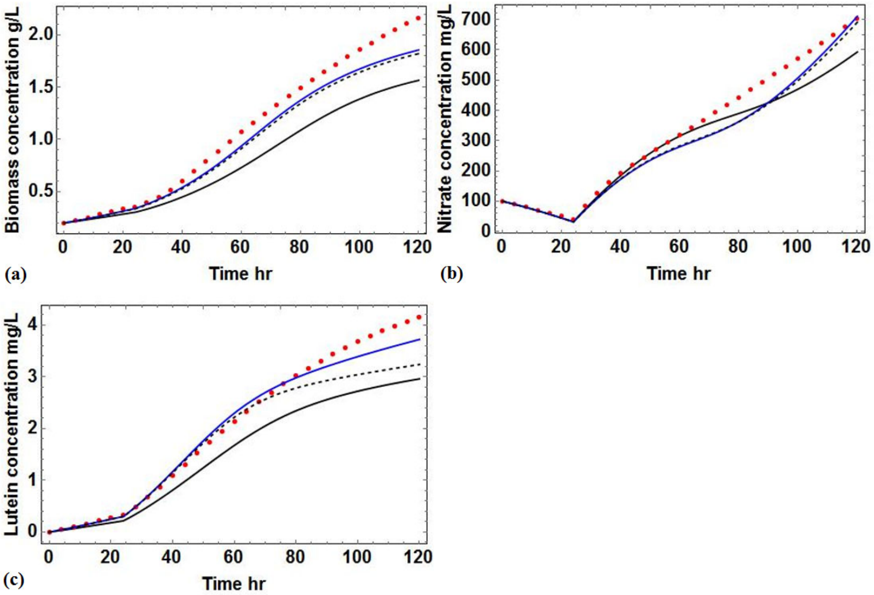


Figure 6: Online model self-calibration result. Lines are the hybrid model prediction results, points are computational experimental data. Black line: initial prediction of the hybrid model (at the beginning of Day 1). Dashed line: prediction of the hybrid model after calibration at the end of Day 1. Blue line: prediction of the hybrid model after calibration at the end of Day 3. (a): biomass concentration; (b): nitrate consumption; (c): lutein production.

4. Conclusion

Overall, based on the current case studies, it is concluded that by adopting the automatic model structure detection framework, it is possible to construct a hybrid model that combines both physical knowledge and insights from the data obtained (through machine learning techniques). Compared to a pure kinetic or a pure data-driven model, the hybrid model does not require deep physical knowledge of the underlying system or a large amount of process data. Based on (computational) experimental verification, it is found that the hybrid model shows significant potential to be used for process optimisation and monitoring, and its self-calibration is easy to implement in an online operating system. Future research will be conducted to further investigate its advantages in process scale-up and reactor design.

It is important to emphasise that although this work conducted model structure identification and parameter estimation simultaneously, in practice this can be executed in sequence depending on the complexity of the kinetic part of the hybrid model. With more available physical information, it is possible to develop more accurate kinetic models that better quantify the process behaviour. The more accurate the kinetic part is, the less complex the data-driven part will be. However, solving a highly complex dynamic MINLP problem is mathematically challenging; thus, developing more effective optimisation algorithms leaves an open gap for the construction and industrial use of large scale hybrid models.

References

1. Harun I, Del Rio-Chanona EA, Wagner JL, Lauersen KJ, Zhang D, Hellgardt K. Photocatalytic Production of Bisabolene from Green Microalgae Mutant: Process Analysis and Kinetic Modeling. *Ind Eng Chem Res* . 2018;57(31):10336-10344. doi:10.1021/acs.iecr.8b02509

2. Xie Y-P, Ho S-H, Chen C-Y, et al. Simultaneous enhancement of CO2 fixation and lutein production with thermo-tolerant Desmodesmus sp. F51 using a repeated fed-batch cultivation strategy. *Biochem Eng J* . 2014;86(7):33-40. doi:10.1016/j.bej.2014.02.015

3. Zhang D, Wan M, del Rio-Chanona EA, et al. Dynamic modelling of Haematococcus pluvialis photoinduction for astaxanthin production in both attached and suspended photobioreactors. *Algal Res* . 2016;13(12):69-78. doi:10.1016/j.algal.2015.11.019

4. Zhang D, Del Rio-Chanona EA, Petsagkourakis P, Wagner J. Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. *Biotechnol Bioeng* . 2019;116(11):2919-2930. doi:10.1002/bit.27120

5. Chen LZ, Nguang SK, Chen XD. *Modelling and Optimization of Biotechnological Processes* . Vol 15. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. doi:10.1007/978-3-540-32493-5

6. Bernard O, Dochain D, Genovesi A, Gouze J-L, Guay M. *Bioprocess Control* . (Dochain D, ed.). London, UK: ISTE; 2008. doi:10.1002/9780470611128

7. Del Rio-Chanona EA, Ahmed NR, Wagner J, Lu Y, Zhang D, Jing K. Comparison of physics-based and data-driven modelling techniques for dynamic optimisation of fed-batch bioprocesses. *Biotechnol Bioeng* . 2019;116(11):2971-2982. doi:10.1002/bit.27131

8. Fouchard S, Pruvost J, Degrenne B, Titica M, Legrand J. Kinetic modeling of light limitation and sulfur deprivation effects in the induction of hydrogen production with Chlamydomonas reinhardtii: Part I. Model development and parameter identification. *Biotechnol Bioeng* . 2009;102(1):232-277. doi:10.1002/bit.22034

9. del Rio-Chanona EA, Liu J, Wagner JL, et al. Dynamic modeling of green algae cultivation in a photobioreactor for sustainable biodiesel production. *Biotechnol Bioeng* . 2018;115(2):359-370. doi:10.1002/bit.26483

10. Jing K, Tang Y, Yao C, del Rio-Chanona EA, Ling X, Zhang D. Overproduction of L-tryptophan via simultaneous feed of glucose and anthranilic acid from recombinant Escherichia coli W3110: Kinetic modeling and process scale-up. *Biotechnol Bioeng* . 2018;115(2):371-381. doi:10.1002/bit.26398

11. do Carmo Nicoletti M, Jain LC, eds. *Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control* . Vol 218. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009. doi:10.1007/978-3-642-01888-6

12. Petsagkourakis P, Sandoval IO, Bradford E, Zhang D, del Rio-Chanona EA. Reinforcement learning for batch bioprocess optimization.*Comput Chem Eng* . 2020;133:106649. doi:10.1016/j.compchemeng.2019.106649

13. Bradford E, Schweidtmann AM, Zhang D, Jing K, del Rio-Chanona EA. Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. *Comput Chem Eng* . 2018;118:143-158. doi:10.1016/j.compchemeng.2018.07.015

14. del Rio-Chanona EA, Wagner JL, Ali H, Fiorelli F, Zhang D, Hellgardt K. Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design. *AIChE J* . 2019;65(3):915-923. doi:10.1002/aic.16473

15. Baughman DR, Liu YA. *Neural Networks in Bioprocessing and Chemical Engineering* . Elsevier; 1995. doi:10.1016/C2009-0-21189-5

16. Oliveira R. Combining first principles modelling and artificial neural networks: a general framework. *Comput Chem Eng* . 2004;28(5):755-766. doi:10.1016/j.compchemeng.2004.02.014

17. von Stosch M, Oliveira R, Peres J, Feyo de Azevedo S. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Comput Chem Eng* . 2014;60:86-101. doi:10.1016/j.compchemeng.2013.08.008

18. Carinhas N, Bernal V, Teixeira AP, Carrondo MJ, Alves PM, Oliveira R. Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products. *BMC Syst Biol* . 2011;5(1):34. doi:10.1186/1752-0509-5-34

19. Teixeira A, Cunha AE, Clemente JJ, et al. Modelling and optimization of a recombinant BHK-21 cultivation process using hybrid grey-box systems. *J Biotechnol* . 2005;118(3):290-303. doi:10.1016/j.jbiotec.2005.04.024

20. Portela RMC, von Stosch M, Oliveira R. Hybrid semiparametric systems for quantitative sequence-activity modeling of synthetic biological parts. *Synth Biol* . 2018;3(1). doi:10.1093/synbio/ysy010

21. Xie Y, Ho S-H, Chen C-NN, et al. Phototrophic cultivation of a thermo-tolerant Desmodesmus sp. for lutein production: Effects of nitrate concentration, light intensity and fed-batch operation.*Bioresour Technol* . 2013;144:435-444. doi:10.1016/j.biortech.2013.06.064

22. del Rio-Chanona EA, Ahmed N rashid, Zhang D, Lu Y, Jing K. Kinetic modeling and process analysis for Desmodesmus sp. lutein photo-production. *AIChE J* . 2017;63(7):2546-2554. doi:10.1002/aic.15667

23. Aguirre A-M, Bassi A. Investigation of biomass concentration, lipid production, and cellulose content in Chlorella vulgaris cultures using response surface methodology. *Biotechnol Bioeng* . 2013;110(8):2114-2122. doi:10.1002/bit.24871

24. Wang J, Wan W. Optimization of fermentative hydrogen production process by response surface methodology. *Int J Hydrogen Energy* . 2008;33(23):6976-6984. doi:10.1016/j.ijhydene.2008.08.051

25. Wang Z, Georgakis C. New Dynamic Response Surface Methodology for Modeling Nonlinear Processes over Semi-infinite Time Horizons. *Ind Eng Chem Res* . 2017;56(38):10770-10782. doi:10.1021/acs.iecr.7b02381

26. Guerra NP. Modeling the batch bacteriocin production system by lactic acid bacteria by using modified three-dimensional Lotka–Volterra equations. *Biochem Eng J* . 2014;88:115-130. doi:10.1016/j.bej.2014.04.010

27. Adesanya VO, Davey MP, Scott SA, Smith AG. Kinetic modelling of growth and storage molecule production in microalgae under mixotrophic and autotrophic conditions. *Bioresour Technol* . 2014;157:293-304. doi:10.1016/j.biortech.2014.01.032

28. Zhang D, Dechatiwongse P, Del-Rio-Chanona EA, Hellgardt K, Maitland GC, Vassiliadis VS. Analysis of the cyanobacterial hydrogen photoproduction process via model identification and process simulation.*Chem Eng Sci* . 2015;128:130-146. doi:10.1016/j.ces.2015.01.059

29. del Rio-Chanona EA, Zhang D, Xie Y, Manirafasha E, Jing K. Dynamic Simulation and Optimization for Arthrospira platensis Growth and C-Phycocyanin Production. *Ind Eng Chem Res* . 2015;54(43):10606-10614. doi:10.1021/acs.iecr.5b03102

30. Yang A. Modeling and Evaluation of CO 2 Supply and Utilization in Algal Ponds. *Ind Eng Chem Res* . 2011;50(19):11181-11192. doi:10.1021/ie200723w

31. Violet L, Loubière K, Rabion A, et al. Stoichio-kinetic model discrimination and parameter identification in continuous microreactors.*Chem Eng Res Des* . 2016;114:39-51. doi:10.1016/j.cherd.2016.07.025

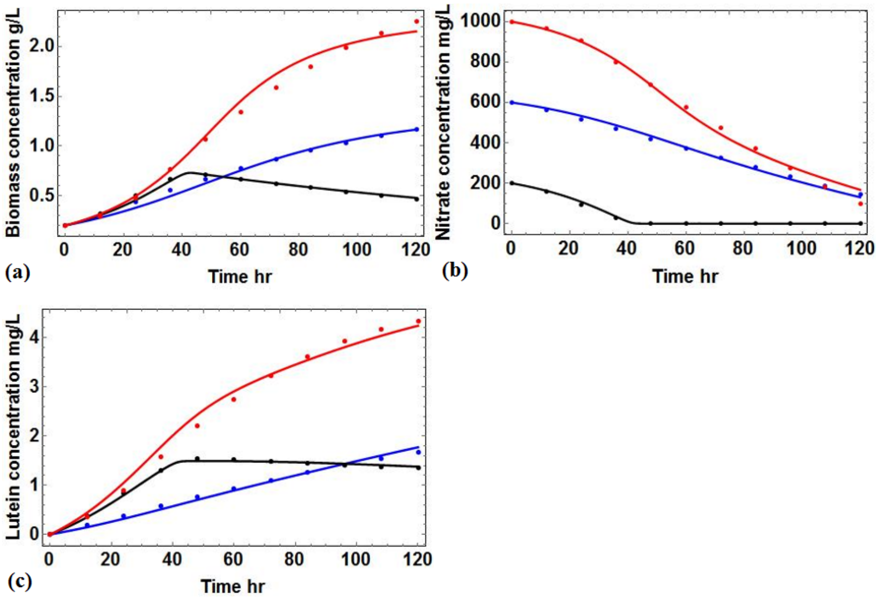
32. Del Rio-Chanona EA, Fiorelli F, Zhang D, Ahmed NR, Jing K, Shah N. An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process. *Biotechnol Bioeng* . 2017;114(11):2518-2527. doi:10.1002/bit.26373

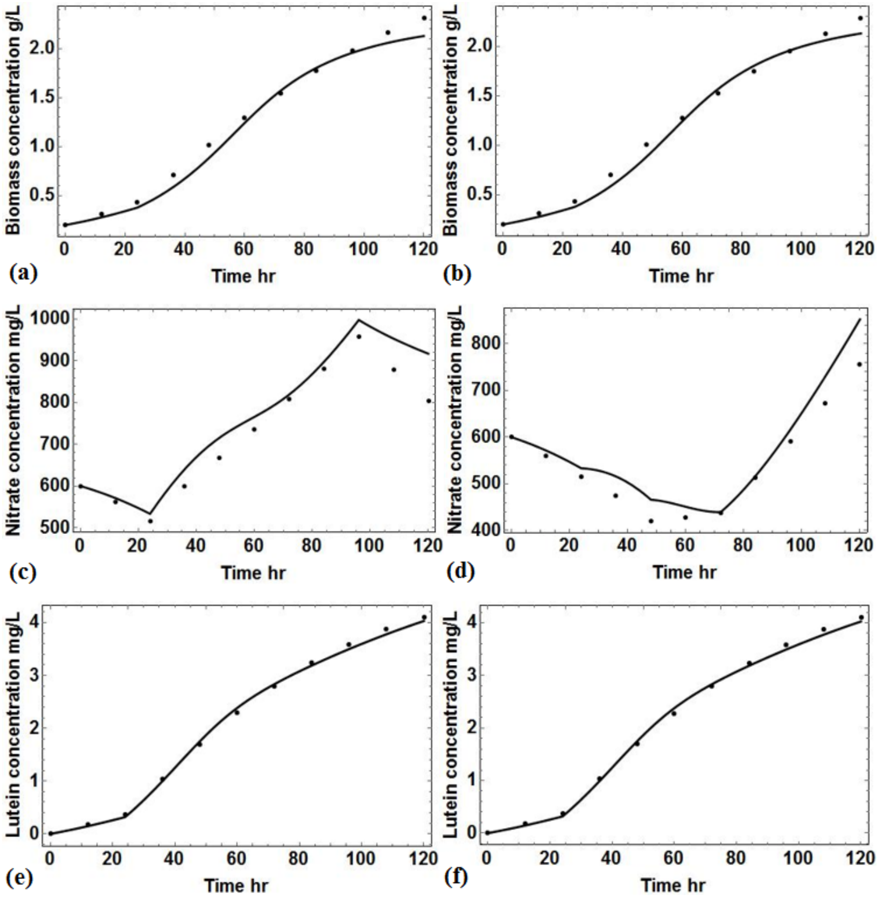
33. Schmidt M, Lipson H. Distilling Free-Form Natural Laws from Experimental Data. *Science (80-)* . 2009;324(5923):81-85. doi:10.1126/science.1165893

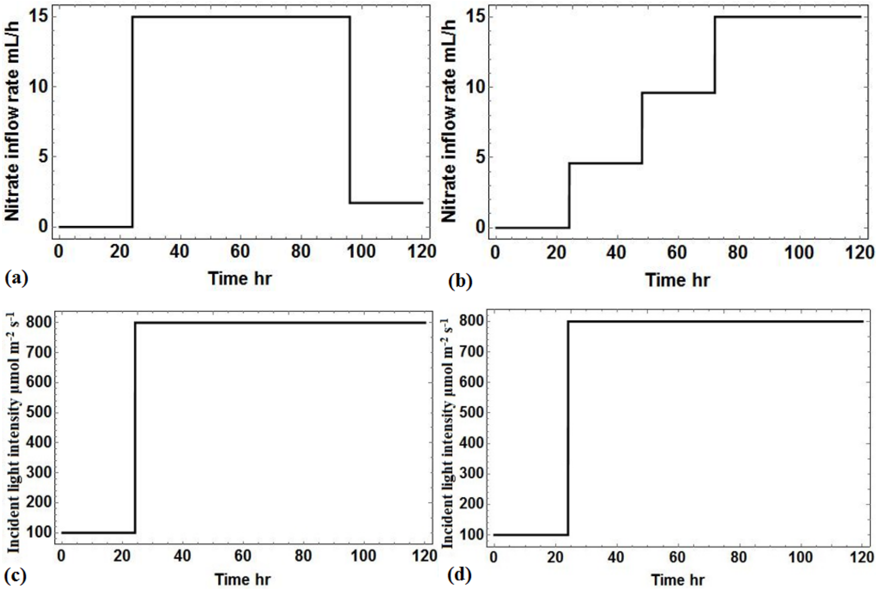
34. Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems.*Proc Natl Acad Sci* . 2016;113(15):3932-3937. doi:10.1073/pnas.1517384113

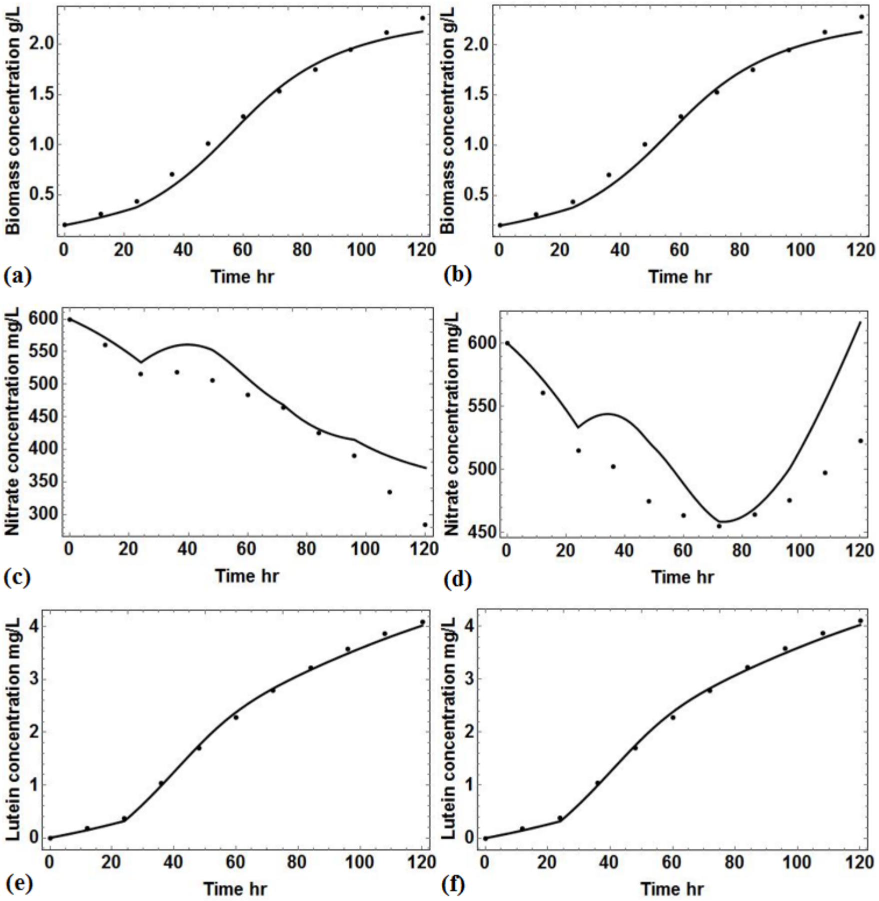
35. del Rio-Chanona EA, Zhang D, Shah N. Sustainable biopolymer synthesis via superstructure and multiobjective optimization.*AIChE J* . 2018;64(1):91-103. doi:10.1002/aic.15877

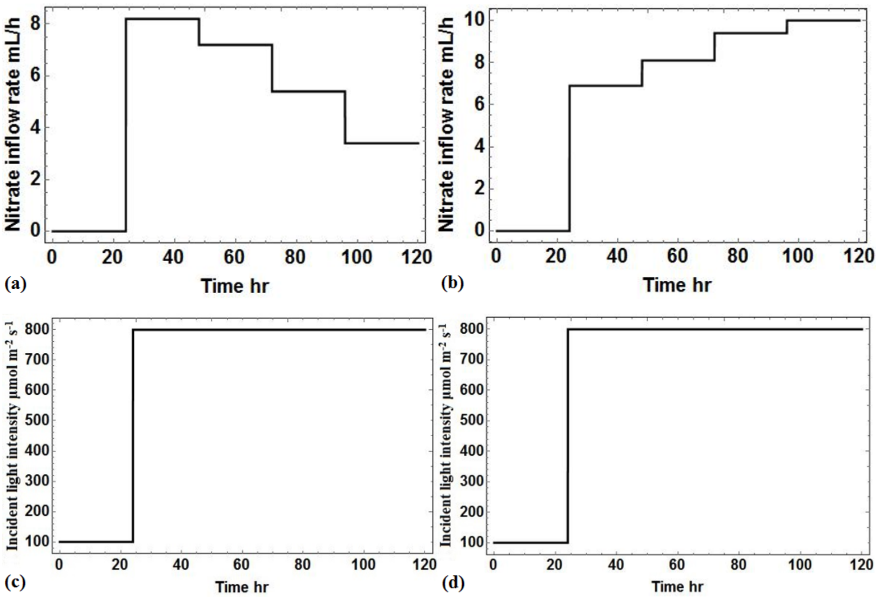
36. Hart WE, Laird C, Watson J-P, Woodruff DL. *Pyomo – Optimization Modeling in Python* . Vol 67. Boston, MA: Springer US; 2012. doi:10.1007/978-1-4614-3226-5

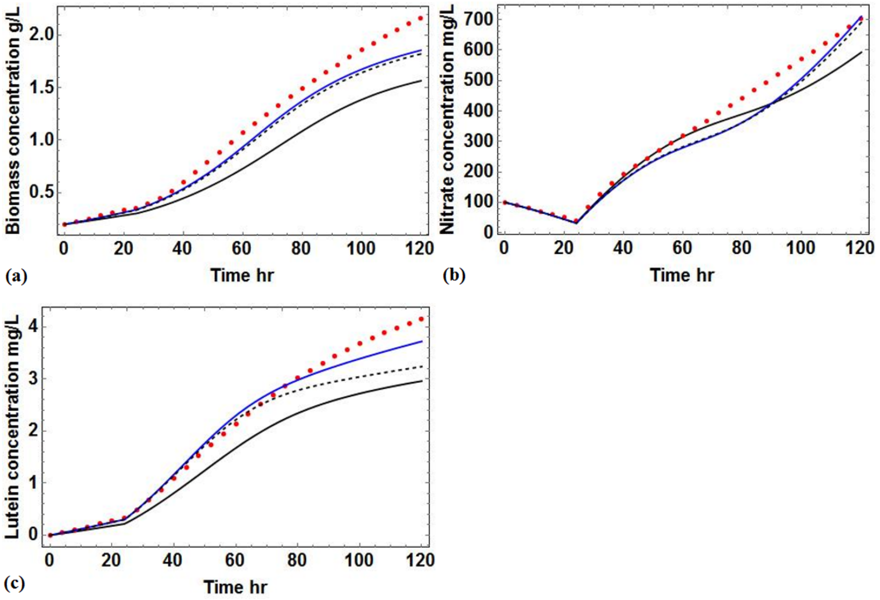





Nitrate inflow rate mL/h
0
5
10
15
0
20
40
60
80
100
120
Time hr
(a)
(b)
Incident light intensity µmol m^{-2} s^{-1}
100
200
300
400
500
600
700
800
(c)
(d)


Biomass concentration g/L
0.5
1.0
1.5
2.0
Time hr
0
20
40
60
80
100
120
(a)
(b)
Nitrate concentration mg/L
300
350
400
450
500
550
600
(c)
(d)
Lutein concentration mg/L
0
1
2
3
4
(e)
(f)


Nitrate inflow rate mL/h
0
2
4
6
8
0
20
40
60
80
100
120
(a)
Time hr
Nitrate inflow rate mL/h
0
2
4
6
8
10
0
20
40
60
80
100
120
(b)
Time hr
Incident light intensity μmol m-2 s-1
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120
(c)
Time hr
Incident light intensity μmol m-2 s-1
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120
(d)
Time hr


Biomass concentration g/L
0.5
1.0
1.5
2.0
0
20
40
60
80
100
120
(a)
Time hr
Nitrate concentration mg/L
0
100
200
300
400
500
600
700
0
20
40
60
80
100
120
(b)
Time hr
Lutein concentration mg/L
0
1
2
3
4
0
20
40
60
80
100
120
(c)
Time hr