Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation

Dongda Zhang | Thomas R. Savage | Bovinille A. Cho

Centre for Process Integration, University of Manchester, The Mill, Manchester, UK

Correspondence
Dongda Zhang, Centre for Process Integration, University of Manchester, The Mill, Sackville Street, Manchester M1 3AL, UK.
Email: dongda.zhang@manchester.ac.uk

Abstract
Integrating physical knowledge and machine learning is a critical aspect of developing industrially focused digital twins for monitoring, optimisation, and design of microalgal and cyanobacterial photo-production processes. However, identifying the correct model structure to quantify the complex biological mechanism poses a severe challenge for the construction of kinetic models, while the lack of data due to the time-consuming experiments greatly impedes applications of most data-driven models. This study proposes the use of an innovative hybrid modelling approach that consists of a simple kinetic model to govern the overall process dynamic trajectory and a data-driven model to estimate mismatch between the kinetic equations and the real process. An advanced automatic model structure identification strategy is adopted to simultaneously identify the most physically probable kinetic model structure and minimum number of data-driven model parameters that can accurately represent multiple data sets over a broad spectrum of process operating conditions. Through this hybrid modelling and automatic structure identification framework, a highly accurate mathematical model was constructed to simulate and optimise an algal lutein production process. Performance of this hybrid model for long-term predictive modelling, optimisation, and online self-calibration is demonstrated and thoroughly discussed, indicating its significant potential for future industrial application.

KEYWORDS
automatic model identification, bioprocess optimisation, fed-batch operation, hybrid modelling, machine learning

1 | INTRODUCTION

Developing sustainable biotechnology for bulk chemicals and high-value products synthesis is an important research theme at this moment. Over the last decades, microalgae and cyanobacteria based photo-production processes have been extensively investigated as they can directly use solar energy and CO₂ to produce commercial materials and renewable fuels (Mata, Martins, & Caetano, 2010; Vavitsas, Fabris, & Vickers, 2018). Enormous effort has been made to construct mathematical models for photo-production systems design, operation, optimisation, and scale-up. To quantify their complex metabolic mechanisms, elaborate kinetic models have been...
developed by embedding new physical understandings (e.g. adding the Aiba model for light intensity effects, and the Beer-Lambert’s law for light attenuation) into classic models such as the Monod and the Droop model (Fouchard, Pruvost, Degrenne, Titica, & Legrand, 2009; Packer et al., 2011; Ryu et al., 2018). However, identifying a correct model structure to describe the physical knowledge is a challenging task, usually with long development times for parameter estimation and model structure selection. This often results in a complex model structure leading to issues with parameter estimation and further use in process optimisation (do Carmo Nicoletti & Jain, 2009). Moreover, as there is still critical biological knowledge yet undetermined, pure kinetic models cannot provide satisfactory accuracy for process long-term prediction, sacrificing their industrial applicability.

On the other hand, machine learning-based data-driven models such as artificial neural networks (ANNs) and Gaussian processes (GPs) have been applied in parallel to avoid the above problems for the predictive modelling and optimal control of photo-production processes (Bradford, Schweidtmann, Zhang, Jing, & del Rio-Chanona, 2018; Liyanaarachchi, Nishshanka, Nimarshana, Ariyadasa, & Attalage, 2020). Although these data-driven models can better capture the complex process dynamics in a specific operating range without the need of prior physical knowledge, they cannot be extrapolated to a wider operating range and are prone to overfitting (Bernard, Dochain, Genovesi, Gouze, & Guay, 2008; do Carmo Nicoletti & Jain, 2009). Most importantly, having a large size (e.g., hundreds to thousands) of data is a prerequisite for data-driven model construction. However, as photo-production processes are slow dynamic systems (e.g., general algae and cyanobacteria doubling time ~20 and 7–12 hr, respectively, compared to bacteria (~0.5 hr) and yeast (~1.5 hr; Bernstein et al., 2016; Liu, Huang, & Che, 2011), it usually takes several weeks or months to complete one experiment for data collection. Hence, acquiring substantial data for photo-production systems modelling is a severe practical obstacle.

In addition, after decades of research, substantial physical information has been discovered for photo-production processes. Thus, pure data-driven models, which are mainly used for systems with little mechanistic information, are not suited for photo-production systems. To this end, hybrid modelling, which aims to integrate prior domain knowledge and machine learning techniques, provides a potentially better simulation strategy. The concept of hybrid modelling was proposed in 1992 (Psichogios & Ungar, 1992) for fermentation process control, and the methodology was developed since the 2000s (Feyo de Azevedo, Dahm, & Oliveira, 1997; Oliveira, 2004). The structure of a hybrid model is flexible depending on the approach to combine data-driven models and mechanistic models (von Stosch, Oliveira, Peres, & Feyo de Azevedo, 2014). In recent years, there have been several studies attempting to apply this strategy into biochemical and general chemical engineering applications (Bangi & Kwon, 2020; Carinhas et al., 2011; Portela, van Stosch, & Oliveira, 2018).

However, all of the previous studies exclusively used ANNs as the data-driven part for hybrid model construction, resulting in two issues. First, embedding ANNs into a kinetic model can impede the numerical calculation accuracy during dynamic parameter estimation. Although it is still feasible to estimate parameters in the kinetic part (Graefe et al., 1999), the optimisation solution quality can be compromised because of the more nonlinear and complex mathematical structure of the hybrid model (Hamilton, Lloyd, & Flores, 2017; von Stosch, 2011). This can be more problematic if several ANNs are embedded into a kinetic model to substitute its original physical parameters.

Secondly, distinct from the kinetic models used in chemical engineering, most kinetic models used in bioprocesses, in particular photo-production processes, are slightly over-fitted (partially non-identifiable, meaning that some parameters have a large confidence interval or high uncertainty) due to the lack of experimental data which is time consuming to obtain (Kiparissides, Koutinas, Kontoravdi, Mantalaris, & Pistikopoulos, 2011; Lee, Ding, Jayaraman, & Kwon, 2018). Fully identifiable kinetic models which remove parameters that are numerically nonidentifiable but have unique physical meanings are often found to have worse fitting results and unsatisfactory predictions for practical use given their oversimplified model structure (Bernard et al., 2008). Hence, using mildly over-fitted kinetic models for bioprocess simulation is inevitable to balance the trade-off between physical insight and mathematical accuracy. Nonetheless, ANNs usually consist of tens or hundreds of parameters, and little effort has been made to systematically identify the suitable ANN structure inside a hybrid model. Thus, integrating an ANN into a kinetic model will aggravate the overfitting risk, jeopardising the hybrid model’s predictive capability.

As a result, this study aims to develop a more flexible hybrid modelling strategy to combine prior physical knowledge with data-driven techniques. In addition, automatic model structure identification will be applied into this strategy to effectively determine the best model structure with the minimum number of data-driven parameters. As hybrid modelling has never been applied to photo-production processes, this study will use microalgal lutein production as a case study to demonstrate the practical use and performance of this hybrid modelling approach.

2 | METHODOLOGY

2.1 | Computational experiment setup

To illustrate the proposed model construction framework and test the efficiency of the hybrid model in process prediction and optimisation, computational experiments were used in this study. Algal biomass growth and lutein synthesis are mainly affected by light intensity and nitrate concentration (Xie et al., 2013). A complex kinetic model proposed previously was used to generate computational experimental data and is presented in Equation (1a)-(1f) (del Rio-Chanona, Ahmed, Zhang, Lu, & Jing, 2017). This model can adequately 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. The reasons to use computational experiments are that this enables the estimation of "best" process performance (i.e. optimal process calculated by the complex process model) which can be adopted as a benchmark to evaluate the hybrid model’s efficiency, as well as being more time efficient to generate data compared to physical experiments.

\[
\frac{dc_N}{dt} = u_m \cdot k(f) \cdot \frac{c_N}{c_N + K_N} \cdot c_X - u_d c_N,
\]

\[
\frac{dc_X}{dt} = -Y_{X/N} \cdot u_m \cdot k(f) \cdot \frac{c_N}{c_N + K_N} \cdot c_X + F_{in} \cdot c_{N,in},
\]

\[
\frac{dc_L}{dt} = h(f) \cdot \frac{c_N}{c_N + K_{NL}} \cdot c_X - k_d \cdot c_L \cdot c_X,
\]

\[
k(f) = \frac{1}{20} \left( \frac{l_0}{l_0 + k_s + \frac{d}{k_L}} + \sum_{n=1}^{9} \frac{2 \cdot l_n}{l_0 + k_s + \frac{d}{k_L}} + \frac{l_{10}}{l_0 + k_s + \frac{d}{k_L}} \right),
\]

\[
h(f) = \frac{k_m}{20} \left( \frac{l_0}{l_0 + k_L + \frac{d}{k_L}} + \sum_{n=1}^{9} \frac{2 \cdot l_n}{l_0 + k_L + \frac{d}{k_L}} + \frac{l_{10}}{l_0 + k_L + \frac{d}{k_L}} \right),
\]

\[
l(f) = l_0 \cdot (e^{-(l \cdot c_X + K_a) \cdot z}) + e^{-(l \cdot c_X + K_a) \cdot (z-1)},
\]

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. \(u_m\) is maximum cell-specific growth rate, \(K_N\) is nitrate half-velocity constant for cell growth, \(u_d\) is cell-specific decay rate, \(Y_{X/N}\) is nitrate yield coefficient, \(K_m\) is maximum lutein synthesis rate constant, \(K_s\) is nitrate half-velocity constant for lutein synthesis, \(k_d\) is lutein degradation rate constant. \(k(f)\) and \(h(f)\) are functions to simulate light effects on cell growth and lutein synthesis, respectively. \(k_s\) and \(k_L\) are light saturation terms for cell growth and lutein synthesis, respectively. \(k_L\) is light inhibition terms for cell growth and lutein synthesis, respectively. \(\tau\) is cell absorption coefficient, \(K_s\) is the bubble scattering coefficient. \(l\) is the distance from light source, \(z\) is the width of the reactor (0.084 m), \(l_0\) is the local light intensity at a distance of \(\frac{l_n}{D}\) \((n = 1, ..., 10)\) m away from the incident light surface area, while \(l_0\) is the incident light intensity. Parameter values can be found in del Rio-Chanona et al. (2017).

This complex model acts as the "true" experimental system. Initially, three computational experiments (Experiments 1–3) were conducted under a broad spectrum of operating conditions from nitrate-limiting and photo-limitation to nitrate-excessive and photo-inhibition as listed in Table 1. Samples were collected once per 12 hr and data generated in these experiments was used to construct the hybrid model. In total, 3 data sets were produced with 11 data points in each data set. This is to mimic real experimental implementations and ensure that the computational experiments are as realistic as possible (Xie et al., 2013). Once constructed, the hybrid model was first exploited to predict and optimise several fed-batch processes (offline in Table 1) and was subsequently verified via computational experiments and compared against the best process performance. Finally, the hybrid model’s self-calibration efficiency was also evaluated through an online experiment (online in Table 1).

### 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 Equation (2), the principle of a hybrid model is to quantify the underlying bioprocess behaviour by using a kinetic model \(K(S)\) to tackle the known dynamics (physical knowledge) and a data-driven model \(D(S)\) to account for the unknown dynamics. Depending on the amount of known physical knowledge and process data, the structure of \(K(S)\) and \(D(S)\) can be flexible.

\[
\frac{dS}{dt} = K(S) + D(S),
\]

\[
D(S) = q_i S_i + \sum_{j=1}^{M} a_{ij} S_j S_p,
\]

where \(S\) is the state variable vector \(S = (X, N, P)\) and \(X, N, P\) represent the concentrations of biomass, nutrient, and product respectively. \(K(S)\) and \(D(S)\) are the kinetic model and the data-driven model, respectively. \(S_i\) is a state variable, \(M\) is the total number of

| **TABLE 1** Operating conditions of computational experiments |
|-------------------|------------------|-----------------|------------------|------------------|
|                  | Exp. 1 | Exp. 2 | Exp. 3 | Offline | Online |
| Initial biomass concentration (g/L) | 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) | 600    | 200    | 1000   | 600    | 100    |
| Initial lutein production (mg/L) | 0.0    | 0.0    | 0.0    | 0.0    | 0.0    |
| Operating time (hr) | 120    | 120    | 120    | 120    | 120    |

Note: Experiments 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.
state and control variables, \( q_i \) and \( a_i \) 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, the Droop model, and their simple derivatives can be directly adopted. 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 process behaviour has been described by the kinetic model and the process. The nonlinearity of this mismatch is greatly reduced compared to the original bioprocess, as the general process behaviour has been described by the kinetic model and the process. Thus, classic kinetic models are not included here. This aims to keep a simple model structure specific to algal photo-attenuation, and nitrate supply are considered when building \( K(S) \). Other factors such as photo-inhibition, biomass decay, and lutein self-degradation which have been used to construct complex kinetic models are not included here. This aims to keep a simple model structure and reduce the amount of required knowledge. These factors are characterised as the model-plant mismatch and are quantified by the data-driven model. The kinetic part of the hybrid model is presented in Equation (3a)–(3d). It is worth noticing that specific to algal photo-production, 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 Equation (3b)), while others tried to establish a direct interaction between these two (see Equation 3c). Adesanya, Davey, Scott, & Smith, 2014; Yang, 2011; Zhang et al., 2015). Similarly, it is still not clear how significantly light affects lutein synthesis. Thus, Equations (3d) and (3e) are proposed in this study based on different hypotheses (no/weak effect and strong effect, respectively). The final kinetic model structure will be determined via the automatic model identification strategy.

\[
\begin{align*}
\frac{dX}{dt} & = \frac{u_0}{c_N + K_N} \cdot \frac{b_0 e^{-C X}}{b_0 e^{-C X} + k_s} \cdot c_X, \quad (3a) \\
\frac{dC}{dt} & = -\frac{Y_{K/X} \cdot u_0}{c_N + K_N} \cdot \frac{b_0 e^{-C X}}{b_0 e^{-C X} + k_s} \cdot c_X + F_{in} \cdot c_{N,pr}, \quad (3b) \\
\frac{dC}{dt} & = -\frac{Y_{K/X} \cdot u_0}{c_N + K_N} \cdot \frac{b_0 e^{-C X}}{b_0 e^{-C X} + k_s} \cdot c_X + F_{in} \cdot c_{N,pr}, \quad (3c) \\
\frac{dC}{dt} & = \frac{Y_{K/X} \cdot u_0}{c_N + K_N} \cdot \frac{b_0 e^{-C X}}{b_0 e^{-C X} + k_s} \cdot c_X + \tau. \quad (3e)
\end{align*}
\]

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

2.3 Automatic model structure identification

As discussed before, the two challenges for hybrid model construction are identifying the physically correct model structure of the kinetic part and preventing the overfitting of the data-driven part. To address these two challenges simultaneously, an automatic model structure detection strategy which comprises of model reformulation and sparse identification of nonlinear dynamics (SINDy, a mathematical algorithm that aims to identify the minimum number of mathematical terms which can accurately regress a given data set) is proposed in this study. The original idea of this strategy was developed to reveal natural laws based on experimental data using basic mathematical operators without any prior knowledge (Schmidt & Lipson, 2009). This idea was then applied to nonlinear dynamic systems in 2016 to identify the governing equations for fluid dynamics (Brunton, Proctor, & Kutz, 2016). The current study aims to modify this strategy and adopt it into photo-production bioprocess hybrid modelling. More mathematical theory background of model structure detection and sparse optimisation can be found in (Bhadriraj, Narasingam, & Kwon, 2019). Detailed implementation of this strategy is 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 Equation (4a)-(4c). By adding two equality constraints equation (4d)-(4e), different kinetic equations for nitrate uptake and lutein production are merged into one expression shown as Equation (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. \( R_{c_{u1}} \) and \( R_{c_{u2}} \)) is active, with the other being inactive as its binary variable is 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.

\[
\begin{align*}
\frac{dc_N}{dt} &= u_0 \cdot \frac{c_N}{c_N + K_N} \cdot b_0 e^{-t - c_N - 1} + k_s + b_{10} \cdot a_{10} \cdot c_1 \\
&+ \sum_{j=1}^{4} b_{1j} \cdot a_{1j} \cdot c_1 \cdot c_i 
\end{align*}
\]
\( (4a) \)

\[
\begin{align*}
\frac{dc_N}{dt} &= b_{N1} \cdot c_{N1} + b_{N2} \cdot c_{N2} + F_n \cdot c_{N1n} + b_{20} \cdot a_{20} \cdot c_2 \\
&+ \sum_{j=1}^{4} b_{2j} \cdot a_{2j} \cdot c_2 \cdot c_i 
\end{align*}
\]
\( (4b) \)

\[
\begin{align*}
\frac{dc_l}{dt} &= b_{L1} \cdot c_{L1} + b_{L2} \cdot c_{L2} + b_{30} \cdot a_{30} \cdot c_3 + \sum_{j=1}^{4} b_{3j} \cdot a_{3j} \cdot c_3 \cdot c_i 
\end{align*}
\]
\( (4c) \)

\[
\begin{align*}
b_{N1} + b_{N2} &= 1, \\
b_{L1} + b_{L2} &= 1 
\end{align*}
\]
\( (4d), (4e) \)

where \( b_{Ni} \) and \( b_{Lj} \) (\( i = 1,2 \)) are binary variables (i.e. either 0 or 1) for the kinetic model, \( b_j \) (\( i = 1,2,3,j = 0,1,2,3 \)) are binary variables for the data-driven model, \( a_j \) (\( i = 1,2,3,j = 0,1,2,3,4 \)) 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. \( c_{N1} \) and \( c_{N2} \) are the first term on the right-hand side of Equations (3b) and (3c), respectively. \( c_{L1} \) and \( c_{L2} \) are the first term on the right-hand side of Equation (3d) and (3e), 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 three experimental data sets. Thus, the objective function for parameter estimation, Equation (5), 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 via the established dynamic parameter estimation algorithm (del Rio-Chanona et al., 2015):

\[
\begin{align*}
\min F &= \sum_{k=1}^{m} \sum_{p=1}^{m} \sum_{i=1}^{3} (g_{i,p,k,E} - g_{i,p,k,M})^2 \cdot W_{i,k,p} + W_b \cdot \sum_{j=0}^{4} \sum_{i=1}^{3} b_{ij} 
\end{align*}
\]
\( (5) \)

where \( g_{i,p,k,E} \) and \( g_{i,p,k,M} \) are experimental and model-simulated value for concentration of state \( S_i \) at time step \( p \in m \) in the \( k \)th data set (total number of data sets is \( n \)), respectively, \( w_{i,k,p} \) and \( w_b \) are the weight for each data point and the sum of binary variables, respectively.

In this study, 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 established dynamic parameter estimation method. Conventionally, an MINLP problem is relaxed via introducing new parameters and extra linear constraints to reduce non-linearity of the original problem (del Rio-Chanona, Zhang, & Shah, 2018). For example, converting Equations (4a)–(6a) where new continuous parameters \( b_{ij} \) are introduced to substitute the bilinear term \( b_{ij} \cdot a_{ij} \) and satisfy the constraints equation (6b) and (6c):

\[
\begin{align*}
\frac{dc_N}{dt} &= u_0 \cdot \frac{c_N}{c_N + K_N} \cdot b_0 e^{-t - c_N - 1} + k_s + b_{10} \cdot a_{10} \cdot c_1 \\
&+ \sum_{j=1}^{4} b_{ij} \cdot a_{ij} \cdot c_1 \cdot c_i 
\end{align*}
\]
\( (6a) \)

\[
\begin{align*}
a_{ij} - a_{ij}^L \cdot (1 - b_{ij}) \leq b_{ij} \leq a_{ij} - a_{ij}^L \cdot (1 - b_{ij}) 
\end{align*}
\]
\( (6b) \)

\[
\begin{align*}
a_{ij}^L \cdot b_{ij} \leq b_{ij} \leq a_{ij}^L \cdot b_{ij} 
\end{align*}
\]
\( (6c) \)

where \( a_{ij}^L \) and \( a_{ij}^U \) are the lower and upper bound of \( a_{ij} \), respectively.

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

\[
\begin{align*}
0 \leq b_{ij} \leq 1. 
\end{align*}
\]
\( (7a) \)

\[
\begin{align*}
b_{ij} \cdot (1 - b_{ij}) \leq 0. 
\end{align*}
\]
\( (7b) \)

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 (Hart, Laird, Watson, & Woodruff, 2012). Confidence intervals were also calculated during parameter estimation, and dynamic optimisation was performed in Python following the same procedure. Model uncertainty analysis was carried out using the standard package in the commercial software Wolfram Mathematica 12.0.

3 | RESULTS AND DISCUSSION

3.1 | Result of hybrid model construction

Through automatic model structure identification, the hybrid model is identified as Equation (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. Figure 1 shows the modelling fitting result. 

\[
d\frac{c_x}{dt} = u_0 \cdot \frac{c_N}{c_N + K_N} \cdot \frac{b_0 e^{-c_K c_x^2}}{b_0 e^{-c_K c_x^2} + K_s} \cdot c_X + a_{10} \cdot c_1 + a_{14} \cdot c_1 \cdot c_A.
\]  

(8a)

\[
d\frac{c_N}{dt} = -Y_{N/X} \cdot u_0 \cdot \frac{c_N}{c_N + K_N} \cdot \frac{b_0 e^{-c_K c_x^2}}{b_0 e^{-c_K c_x^2} + K_s} \cdot c_X + F_{in} \cdot c_{N,in} + a_{20} \cdot c_2.
\]  

(8b)

\[
d\frac{c_L}{dt} = Y_{L/X} \cdot u_0 \cdot \frac{c_N}{c_N + K_N} \cdot \frac{b_0 e^{-c_K c_x^2}}{b_0 e^{-c_K c_x^2} + K_{sl}} \cdot c_X + a_{31} \cdot c_3 \cdot c_5
\]

+ a_{33} \cdot c_1 \cdot c_B.  

(8c)

The 95% confidence intervals of kinetic parameters are listed in Table 2. Confidence intervals of data-driven parameters are calculated for model uncertainty prediction, however, they are not listed in the table as they have no physical meaning (not providing any physical insight to the studied system) and it is uncommon to calculate parameter confidence intervals for a data-driven model such as ANN. It is worth mentioning that most biochemical kinetic models are lumped models and their formulation changes depending on different assumptions (e.g. Monod model, Droop model). Thus, values of the same kinetic parameter may not be fully comparative between different models. The confidence interval of kinetic parameters is to indicate identifiability of current hybrid model. From the table, it is observed that most parameters are identifiable (i.e. narrow confidence interval). However, \(k_{sl}\), the light saturation term for lutein synthesis, is practically non-identifiable given its high uncertainty. This further indicates that due to the sophisticated process mechanism and practical challenge in generating substantial data for photo-production processes, it is advantageous to build hybrid models that consist of a simple kinetic part and a simple data-driven part. Using complex data-driven models or elaborate kinetic models to simulate a photo-production process will inevitably exacerbate overfitting.

From Figure 1, it is seen that the hybrid model can accurately fit all the three data sets collected over a broad spectrum of

| Parameter | Value       | Parameter | Value       |
|-----------|-------------|-----------|-------------|
| \(\omega_0\) (hr\(^{-1}\)) | 0.045 ± 0.00142 | \(Y_{N/X}\) (mg/g) | 315.2 ± 15.8 |
| \(K_N\) (mg/L) | 6.30 ± 2.798 | \(Y_{L/X}\) (mg/g) | 4.44 ± 1.839 |
| \(K_s\) (\(\mu\)mol·m\(^{-2}\)·s\(^{-1}\)) | 16.7 ± 0.821 | \(k_{sl}\) (\(\mu\)mol·m\(^{-2}\)·s\(^{-1}\)) | 155.6 ± 512.4 |
| \(\tau\) (m²/g) | 0.031 ± 0.00201 | \(a_{10}\) | -0.0067 |
| \(a_{14}\) | 0.0024 | \(a_{20}\) | 0.216 |
| \(a_{31}\) | 0.012 | \(a_{33}\) | -0.0055 |

**TABLE 2** Parameter values of the current hybrid model

**FIGURE 1** Curve fitting result of the hybrid model and pure kinetic model. Solid lines are the hybrid model fitting results, dashed lines are the pure kinetic model fitting results, and points are computational experimental data. Bands are the model uncertainty. Blue, black, and red colours refer to Experiments 1, 2, and 3, respectively. (a) Biomass concentration. (b) Nitrate consumption. (c) Lutein production [Color figure can be viewed at wileyonlinelibrary.com]
operating conditions (e.g., substrate limiting to substrate excessive, low light intensity to high light intensity). Compared to the pure kinetic model which has an average fitting error of 6.1%, 8.5%, and 7.3% for concentrations of biomass, nitrate, and lutein, respectively, the fitting error of the hybrid model is only 1.7%, 4.6%, and 2.9% for concentrations of biomass, nitrate, and lutein, respectively. In addition, the total number of binary variables assigned to the data-driven model is 15, while only 5 are active after sparse optimisation. The kinetic formulation is also identified more efficiently than traditional model discrimination. Compared to the complex kinetic model shown in Equation (1a)–(1d), the current hybrid model adequately fits all the data sets with a simpler model structure. As the parameter \( k_{sl} \) accounting for lutein production is practically non-identifiable, it is found that model uncertainty is slightly higher when predicting lutein production. Overall, it is concluded that integrating automatic model structure identification into hybrid model construction is advantageous.

### 3.2 Model validation and offline optimisation

To further validate accuracy of this hybrid model, it is first used to predict optimal control actions for a long-term fed-batch lutein production process, which is subsequently verified experimentally. The objective function is to maximise the final lutein production over a 5-day operation system in which incident light intensity (100–800 \( \mu \)mol·m\(^{-2}\)·s\(^{-1}\)) and nitrate inflow rate (0–17.5 ml/hr 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 predicted by the hybrid model, they were implemented and verified by the computational experiment. In addition, the original complex kinetic model was also used to search for the theoretical maximum lutein production in this fed-batch process. This is used as a benchmark to test the hybrid model’s efficiency. Initial operating conditions of this process are listed in Table 1.

Figures 2 and 3 show the model prediction result and experimental verification results. From Figure 2a,c,e, the hybrid model’s
prediction is verified to be accurate. From Figure 2b,df, it is also observed that the hybrid model can well predict the theoretical best process behaviour. The prediction error of the hybrid model is 3.4%, 1.9%, and 3.1% for concentrations of biomass, nitrate and lutein, respectively. The maximum model prediction uncertainty is 8.7% (final lutein production). When comparing the two processes, the hybrid model is found to result in the same lutein production as the theoretical best process, indicating its high performance in offline optimisation. The two processes (identified by the hybrid model: Figures 2a,e, and 3c); theoretical best process: Figure 2b,f,d follows a similar trend for biomass growth, lutein production, and control actions of incident light intensity. However, culture nitrate concentration (Figure 2b vs. 2d) and nitrate inflow rate (Figure 3a vs. 3b) 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 (400–900 mg/L). It also suggests that multiple optimal solutions exist for this offline optimisation problem. The uncertainty of the high model’s prediction gradually increases along with time as expected because of error propagation. However, uncertainty remains low throughout the entire course.

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/hr; the second being that final culture nitrate concentration must be lower than 600 mg/L; and the third being that total lutein production must be no lower than 4.0 mg/L. 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 experimental verification as shown in Figure 4, it is concluded that, once again, the hybrid model well predicted the dynamics of its optimised process. It also predicted well concentrations of biomass (prediction error 3.4%) and lutein (prediction error 3.1%) in the theoretical best process (with a maximum prediction uncertainty of 8.7% on final lutein production). Nonetheless, its prediction of nitrate concentration in the theoretical best process (Figure 4d) is found to have large deviations (prediction error 7.1%). In addition, its uncertainty in nitrate concentration (Figure 4c,d) increases significantly (maximum model uncertainty 12.7% for final nitrate concentration) during the later stage of operation, possibly attributed to the frequent changes of nitrate inflow rate. 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.

3.3 | Online calibration for state estimation

In practice, new measurement of state variables (offline data) can be collected through regular sampling and offline analysis during an ongoing process. However, it is not easy to directly use this data to 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
update its large number of parameters. Nonetheless, offline data collected from an ongoing process is usually limited, hence not able to satisfy this prerequisite.

Another great advantage of hybrid modelling is that it can solve both issues relatively easily. 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., five parameters in this case), thus offline data (usually sampled once per 4 hr in an industrial process, hence 6 data points/day) collected from a process can be used to update the model’s accuracy on a daily basis. Parameter re-estimation of the current hybrid model takes seconds to complete, having minimum impact on delaying process decision-making or estate estimation. As a result, online calibrating a hybrid model is feasible and 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 such 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 Figure 5 when only the initial operating conditions and pre-determined control actions are given. It is seen that six large mismatch exists at a later stage of the process. However, after model recalibration at the end of Day 1 (using 6 new data points collected in Day 1), it is found that the model’s prediction is improved significantly, particularly for biomass (average prediction error reduced from 11.7% to 1.5%) and lutein concentrations (average error reduced from 14.3% to 2.1%). 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 (average prediction error ~2.7%). This is because the

**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) Concentrations of biomass, lutein, and nitrate in the fed-batch process optimised by the hybrid model. (d) Nitrate concentration in the theoretical best process. (e, f) Optimal nitrate inflow rate identified by the hybrid model and the theoretical best process, respectively. As the optimal control actions of light intensity are found to be the same as Figure 3c,d, they are not repeated here.
kinetic part of the model comprises a 15% error in its parameters and has never been calibrated. In practice, if enough new data points can be sufficiently accumulated from an ongoing process, they can be used to re-estimate values of all the parameters in the hybrid model to further improve model accuracy.

4 | CONCLUSION

Overall, based on the current case studies, it is concluded that by adopting the automatic model structure detection strategy, it is possible to construct a flexible hybrid model that combines both prior 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 understanding 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 study 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, one can first identify the kinetic model parameters if the kinetic part is already complex (e.g., more available physical knowledge), and then use the SiNDy algorithm to determine the data-driven part formulation and parameters. In addition, a hybrid model should be mainly considered when there is prior domain knowledge of the underlying process. The complexity of the kinetic part in a hybrid model is case by case, and relies on the modeller’s experience. Based on the current work, we would recommend that one should only use the kinetic part to simulate primary process mechanisms, and use the data-driven part to account for secondary process mechanisms. This may help with constructing an accurate yet simple hybrid model structure.

ORCID

Dongda Zhang http://orcid.org/0000-0001-5956-4618
Thomas R. Savage http://orcid.org/0000-0001-8715-8369

REFERENCES

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

Aguirre, A.-M., & Bassi, A. (2013). Investigation of biomass concentration, lipid production, and cellulose content in Chlorella vulgaris cultures using response surface methodology. Biotechnology and Bioengineering, 110(8), 2114–2122. https://doi.org/10.1002/bit.24871
Graefe, J., Bogaerts, P., Castillo, J., Cherlet, M., Wérenne, J., Marenbach, P., & Hanus, R. (1999). A new training method for hybrid models of bioprocesses. Bioprocess Engineering, 21(5), 423. https://doi.org/10.1002/biop.22034

Guerra, N. P. (2014). Modeling the batch bacteriocin production system by lactic acid bacteria by using modified three-dimensional Lotka-Volterra equations. Biochemical Engineering Journal, 88, 115–130. https://doi.org/10.1016/j.bej.2014.04.010

Hamilton, F., Lloyd, A. L., & Flores, K. B. (2017). Hybrid modeling and prediction of dynamical systems. PLOS Computational Biology, 13(7), e1005655. https://doi.org/10.1371/journal.pcbi.1005655

Hart, W. E., Laird, C., Watson, J.-P., & Woodruff, D. L. (2012). Pyomo—Optimization modeling in python (67). Boston, MA: Springer. https://doi.org/10.1007/978-1-4614-3226-5

Kiparisides, A., Koutinas, M., Kontoravdi, C., Mantalaris, A., & Pistikopoulos, E. N. (2011). ‘Closing the loop’ in biological systems modeling—From the in silico to the in vitro. Automatica, 47(6), 1147–1155. https://doi.org/10.1016/j.automatica.2011.01.013

Lee, D., Ding, Y., Jayaraman, A., & Kwon, J. (2018). Mathematical modeling and parameter estimation of intracellular signaling pathway: Application to LPS-induced NF-kB activation and TNFα production in macrophages. Processes, 6(3), 21. https://doi.org/10.3390/pr6030021

Liu, J., Huang, J., & Che, F. (2011). Microalgae as feedstocks for biodiesel production. Biodiesel—Feedstocks and processing technologies. InTech. https://doi.org/10.5772/25600

Liyanaarachchi, V. C., Nishshanka, G. K. S. H., Nimarshana, P. H. V., Arijiyadasa, T. U., & Attalage, R. A. (2020). Development of an artificial neural network model to simulate the growth of microalgae Chlorella vulgaris incorporating the effect of micronutrients. Journal of Biotechnology, 312, 44–55. https://doi.org/10.1016/j.jbiotec.2020.02.010

Mata, T. M., Martins, A. a, & Caetano, N. S. (2010). Microalgae for biodiesel production and other applications: A review. Renewable and Sustainable Energy Reviews, 14(1), 217–232. https://doi.org/10.1016/j.rser.2009.07.020

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: A general framework. Computers & Chemical Engineering, 28(5), 755–766. https://doi.org/10.1016/j.compchemeng.2004.02.014

Packer, A., Li, Y., Andersen, T., Hu, Q., Kuang, Y., & Sommerfeld, M. (2011). Growth and neutral lipid synthesis in green microalgae: A mathematical model. Bioresource Technology, 102(1), 111–117. https://doi.org/10.1016/j.biortech.2010.06.029

Portela, R. M. C., von Stosch, M., & Oliveira, R. (2018). Hybrid semiparametric systems for quantitative sequence-activity modeling of synthetic biological parts. Synthetic Biology, 3(1), https://doi.org/10.1093/synbio/syy010

Psychogiou, D. C. & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. AIChE Journal, 38(10), 1499–1511. https://doi.org/10.1002/aic.690381003

Ryu, K. H., Sung, M.-G., Kim, B., Heo, S., Chang, Y. K., & Lee, J. H. (2018). A mathematical model of intracellular behavior of microalgae for predicting growth and intracellular components syntheses under nutrient-replete and -deplete conditions. Biotechnology and Bioengineering, 115(10), 2441–2455. https://doi.org/10.1002/bit.26744

Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. Science, 324(5923), 81–85. https://doi.org/10.1126/science.1158593

von Stosch, M. (2011). Novel strategies for process control based on hybrid semi-parametric mathematical systems. University of Porto.

Vavitsas, K., Fabris, M., & Vickers, C. (2018). Terpenoid metabolic engineering in photosynthetic microorganisms. Genes, 9(11), 520. https://doi.org/10.3390 Genes9110520

von Stosch, M., Oliveira, R., Peres, J., & Feyo de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Computers & Chemical Engineering, 60, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008
Wang, J., & Wan, W. (2008). Optimization of fermentative hydrogen production process by response surface methodology. *International Journal of Hydrogen Energy, 33*(23), 6976–6984. https://doi.org/10.1016/j.ijhydene.2008.08.051

Wang, Z., & Georgakis, C. (2017). New dynamic response surface methodology for modeling nonlinear processes over semi-infinite time horizons. *Industrial & Engineering Chemistry Research, 56*(38), 10770–10782. https://doi.org/10.1021/acs.iecr.7b02381

Xie, Y., Ho, S.-H., Chen, C.-N. N., Chen, C.-Y., Ng, I.-S., Jing, K.-J., ... Lu, Y. (2013). Phototrophic cultivation of a thermo-tolerant Desmodesmus sp. for lutein production: Effects of nitrate concentration, light intensity and fed-batch operation. *Bioresource Technology, 144*, 435–444. https://doi.org/10.1016/j.biortech.2013.06.064

Yang, A. (2011). Modeling and evaluation of CO₂ supply and utilization in algal ponds. *Industrial & Engineering Chemistry Research, 50*(19), 11181–11192. https://doi.org/10.1021/ie200723w

Zhang, D., Dechatiwongse, P., Del-Rio-Chanona, E. A., Hellgardt, K., Maitland, G. C., & Vassiliadis, V. S. (2015). Analysis of the cyanobacterial hydrogen photoproduction process via model identification and process simulation. *Chemical Engineering Science, 128*, 130–146. https://doi.org/10.1016/j.ces.2015.01.059

How to cite this article: Zhang D, Savage TR, Anye Cho BA. Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation. *Biotechnology and Bioengineering*. 2020;117:3356–3367. https://doi.org/10.1002/bit.27512