Hybrid physics-based and data-driven modelling for bioprocess online simulation and optimisation

Dongda Zhang[1,2,*], Ehecatl Antonio Del Rio-Chanona[2], Panagiotis Petsagkourakis[1,3], Jonathan Wagner[4]

1: Centre for Process Integration, University of Manchester, The Mill, Sackville Street, Manchester, M1 3AL, UK.

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

3: Centre for Process Systems Engineering, University College London, Torrington Place, London, WC1E 7JE, UK.

4: Department of Chemical Engineering, Loughborough University, Loughborough Leicestershire, LE11 3TU, UK.

*: Corresponding author, email: dongda.zhang@manchester.ac.uk, tel: 44 (0)161 306 5153.

Abstract

Model-based online optimisation has not been widely applied to bioprocesses due to the challenges of modelling complex biological behaviours, low-quality industrial measurements, and lack of visualisation techniques for ongoing processes. This study proposes an innovative hybrid modelling framework which takes advantages of both physics-based and data-driven modelling for bioprocess online monitoring, prediction, and optimisation. The framework initially generates high-quality data by correcting raw process measurements via a physics-based noise filter (a generally available simple kinetic model with high fitting but low predictive performance); then constructs a predictive data-driven model to identify optimal control actions and predict discrete future bioprocess behaviours. Continuous future process trajectories are subsequently visualised by re-fitting the simple kinetic model (soft sensor) using the dada-driven model predicted discrete future data points, enabling the accurate monitoring of ongoing processes at any operating time. This framework was tested to maximise fed-batch microalgal lutein production by combining with different online optimisation schemes and compared against the conventional open-loop optimisation technique. The optimal results using the proposed framework were found to be comparable to the theoretically best production, demonstrating its high predictive and flexible capabilities as well as its potential for industrial application.

Keywords: machine learning; data recalibration; kinetic modelling; bioprocess optimisation; fed-batch operation.

1. Introduction

With the rapid development of digital computing technologies, industrially focused mathematical modelling tools have been extensively applied to chemical engineering systems for process simulation, optimisation, control, and design (Marchetti, François, Faulwasser, & Bonvin, 2016; Voll & Marquardt, 2012; Zhang, del Rio-Chanona, & Shah, 2017). Strong global demand for innovative biotechnologies to sustainably produce energy, food, pharmaceuticals, and platform chemicals (Harun et al., 2018; Jeandet, Vasserot, Chastang, & Courot, 2013; Jing et al., 2018) has opened up great opportunities for computer-aided technologies in various bio-production industries *e.g.* fermentation and photo-production (Jing et al., 2018; Wagner, Lee-Lane, Monaghan, Sharifzadeh, & Hellgardt, 2019). As most bioprocesses rely on microorganisms to synthesise the desired products, the simulation of the complex microbial activities has become a critical task, requiring the use of state-of-the-art process systems engineering tools for bioprocess optimisation and scale-up.

Bio-production systems are traditionally simulated using kinetic models which simplify the large number of metabolic reactions into small sets of differential equations (Bernard, Dochain, Genovesi, Gouze, & Guay, 2008). This approach has been highly successful for modelling fermentation processes. More recently, kinetic models have been integrated into cutting-edge online optimisation and state estimation frameworks to improve the productivity of fed-batch systems for biorenewable production (Ehecatl Antonio del Rio-Chanona, Zhang, & Vassiliadis, 2016). Most of these kinetic models are built upon a few classic models such as the Monod model, the Droop model, and the Logistic model (Vatcheva, de Jong, Bernard, & Mars, 2006). Despite adopting simple mathematical structures and having initially been proposed to describe distinct mechanistic hypotheses, these classical models continue to be widely applied in modern industries due to their good data-fitting abilities across a wide range of bioprocesses. This is attributed to the qualitative similarities in behaviour observed

for different species of microorganisms exposed to similar cultivation conditions, thanks to their delicate metabolic regulation mechanisms.

Despite their excellent data-fitting capability, classic kinetic models possess poor predictive capabilities for complex biological systems, *e.g.* algal photo-production or microbial consortia wastewater treatment (Del Rio‐Chanona, Cong, Bradford, Zhang, & Jing, 2018; Zhang et al., 2015). This is caused by the construction principles of the kinetic models themselves, which *lump hundreds of metabolic pathways into a few parameters which are assembled in a simple model structure*. As intracellular metabolic pathways are strongly dependent on the cultivation conditions, any changes in these conditions also have a substantial impact on the values of the model parameters. Because of this, *model parameters estimated from one specific set of data may no longer apply to the same system operated under a different operating condition, even if the model structure is applicable in both cases* (Adesanya, Davey, Scott, & Smith, 2014; Fouchard, Pruvost, Degrenne, Titica, & Legrand, 2009). To address this challenge, elaborate predictive kinetic models have been proposed, which embed additional parameters into the classic kinetic models to account for specific biochemical mechanisms. However, this approach causes highly complex model structures and introduces issues with parameter estimation, model identifiability, amongst others (Bernard et al., 2008; Zhang et al., 2016). Moreover, the flexibility of these models is greatly compromised as they are designed for one specific system, and the lack of knowledge of the underlying processes hampers the effective construction of predictive kinetic models. Hence, kinetic modelling is not always suitable for complex bioprocess prediction.

In recent years, data-driven modelling has been considered as an alternative to kinetic modelling. Compared to kinetic models, data-driven models contain more parameters and structures for data regression, *enabling the inclusion of distinct process behaviours collected at different operating conditions in a single model. This allows the model to accurately*

interpolate the performance of untested new processes operated over a wide range of operating conditions. Given the large size of accumulated data, models based on machine learning, particularly artificial neural networks, have been adopted to simulate bacterial fermentation and algal photo-production (Dineshkumar *et al.*, 2015; del Rio-Chanona *et al.*, 2017). Other advanced neural networks such as recurrent and convolutional neural networks have also been used to simulate the production of different biochemical (Valdez-Castro, *et al.*, 2003). Moreover, through the use of stochastic optimisation, data-driven models have been adopted to optimise long-term fed-batch processes, causing considerable increases in production of different metabolites and yielding highest intracellular contents of bioproducts reported to date (Ehecatl Antonio del Rio-Chanona, Manirafasha, Zhang, Yue, & Jing, 2016; Dineshkumar et al., 2015). Since 2018, Gaussian Process regression, another technique popularised by the machine learning community, was successfully applied for bioprocess modelling and optimisation and online monitoring (Bradford, Schweidtmann, Zhang, Jing, & del Rio-Chanona, 2018; Tulsyan, Garvin, & Ündey, 2018).

Nonetheless, data-driven models are heavily reliant on the quality and quantity of datasets, currently limiting their application to industrial biosystems. Industrial datasets often contain large measurement errors and systematic noise, and it is not feasible to frequently measure all state variables (*e.g.* biomass concentration) at regular time intervals over the entire course of operation (Baughman & Liu, 1995). Secondly, most data-driven models calculate the values of state variables at fixed, pre-specified time intervals. However, data sampled at a plant are often obtained at different time intervals subject to the availability and efficiency of the analysis equipment, causing additional obstacles (*i.e.* missing information) for data-driven model construction. In addition, as many bioproducts are manufactured periodically, the number of accumulated datasets for a specific operation can be much lower than for a continuously operated chemical plant. This lack of high-quality datasets severely hampers the

application of data-driven models to industrial bioprocesses. Another particular challenge with data-driven models is that they are not based on physical mechanisms and are predominantly used to interpolate steady-state chemical processes (Baughman & Liu, 1995). Therefore, their effectiveness in estimating unknown dynamic processes for biochemical plants has been poorly explored so far.

In order to address the limitations of physics-based and data-driven models and to facilitate their application to industrial bioprocesses, this study proposes an innovative hybrid modelling framework that exploits advantages of both modelling approaches and can be easily integrated into several online optimisation strategies. In specific, algal lutein synthesis is chosen as the case study due to its high complexity and increasing market demand.

2. Methodology

2.1 Principles of the hybrid modelling framework

The innovation of this study is to combine the high data-fitting capability of kinetic models with the interpolation power of data-driven models to form a hybrid modelling framework to simulate generic industrial bioprocesses. Given that data-driven models require high quality data, which are rarely available for bioprocesses, it is sensible to initially use a generally available simple kinetic model, based on fundamental biochemical theories, to reduce noise and inconsistencies of *individual* industrial raw datasets. The main reason for selecting a classic kinetic model as the noise filter is that these models have become well established and provide good fits to most existing bioprocesses (*for single datasets*). In addition, kinetic models can be used to generate data in continuous time to complete missing data points from plant records required for data-driven model construction. At the same time, the simple structure of these models avoids the numerical issues associated with parameter estimations of more elaborate predictive kinetic models, which are difficult to construct and do not currently exist for most bioprocesses.

Once high-quality datasets have been obtained from the kinetic model, they can be combined *together* to construct a data-driven model which is capable of simulating process dynamics over a broad operational range. The key reason for developing this new data-driven model for process prediction instead of simply applying the initial classic kinetic model is that, as will be shown in Section 3.1.1, the kinetic model parameters values are substantially different for each specific experimental dataset. Thus, parameters estimated from one dataset cannot be used to calculate the process dynamics under another cultivation condition. This means that the classic or simple kinetic model does not qualify as a predictive model, since *it is unlikely to identify a single set of parameter values that allows the kinetic model to fit well multiple datasets obtained under different operating conditions*.

It is worth emphasising that this conclusion also holds for many recently developed elaborate kinetic models, whose predictive powers have been found to remain restricted to narrow operational ranges even after embedding additional parameters and modifying the original model structures (Adesanya et al., 2014; Fouchard et al., 2009). Although several online parameter estimation techniques can be used to synchronise parameter values, they are mainly used for tuning purpose and therefore not applicable to the current case. Based on these observations, *kinetic models do not appear to be well suited for the estimation of new process dynamics. Instead, it is vital to select a model which can accurately interpolate process data over a broad operational spectrum* (*e.g.* data-driven models).

Once constructed, the data-driven model can be used to predict future process behaviours and identify optimal control actions by exploiting stochastic optimisation. The predictions from the data-driven model are then fed back to the classic kinetic model (acting as a soft sensor) to re-estimate the kinetic model parameters and forecast the full continuous trajectories of future processes, facilitating the visualisation and monitoring of the ongoing plant. The reason for using the kinetic model instead of the data-driven model as the soft sensor is that

data-driven models can only estimate discrete process points, whereas kinetic models can predict the continuous process behaviour over the entire operation cycle.

2.2 Procedures to construct and execute the framework

The hybrid modelling framework consists of three levels as exhibited in Fig. 1: a bottom-level (Level 1) where raw data is collected from the process plant (*i.e.* the real system), a middle-level (Level 2) at which a simple kinetic model is adopted to correct and generate missing data points, and a top-level (Level 3) where a data-driven model is exercised to incorporate different process behaviours and predict optimal control strategies. Procedures to construct and apply this framework are explained in detail in the following sections. This framework incorporates the advantages of both kinetic and data-driven modelling strategies, whilst overcoming the limitations of each individual approach. The kinetic model plays a vital role in generating high quality data and visualising future process behaviours. Given its knowledge-based nature, it is superior to other statistic regression models. Meanwhile, the data-driven model takes advantage of the filtered data and is used to estimate the optimal operating conditions for the process. To illustrate the efficiency of this framework, a case study on the optimisation of fed-batch microalgal lutein production is also presented here.

2.3 Construction of the hybrid modelling framework

2.3.1 Setup of computational experiments

Lutein is a primary carotenoid and a high-value bioproduct that has shown great potential in pharmaceutical and food industries. Following the discovery of a lutein-producing thermo-tolerant microalgae strain, *Desmodesmus* sp. F51, which can achieve 10 times higher intracellular concentrations than the traditional lutein producing plant, marigold flowers (Xie et al., 2013), there has been an increasing economic drive to use microalgae for industrial lutein production. Previous works have shown that lutein synthesis is primarily dependent on light intensity and nitrate concentration (Xie *et al.*, 2013; del Rio-Chanona *et al.*, 2017).

Whilst nitrates are essential for lutein synthesis, lutein production is suppressed at high nitrate concentrations. A different behaviour is observed for light intensity; low light intensities favour lutein accumulation but stifle biomass growth. Therefore, lutein accumulation in the cells has to be balanced against biomass growth to achieve the best overall productivities. To complicate matters more, light absorption by the microalgae cells causes light attenuation within the system, resulting in non-uniform local light intensities inside the bioreactor. In practice, incident light intensities are often kept constant during a fed-batch process and lutein synthesis is exclusively controlled by regulating the inflow rate of nitrate.

A complex mechanistic model was previously designed to simulate the effects of light intensity, nitrate concentration and light attenuation on biomass growth and lutein production (del Rio-Chanona *et al.*, 2017) and is presented in Eq. (1a)-(1f). To test the performance of the newly proposed hybrid modelling framework, computational experiments were used in this work as the initial investigation. The mechanistic model was used to generate three computational datasets at different operating conditions (limited, medium, and excessive nitrate concentration, respectively). Each dataset was modelled for a total duration of 7 days, with a constant incident light intensity of 750 μmol m^{-2} s^{-1}. After 48 hours, a 0.2M nitrate feed was started at different flowrates for each computational experiment. Detailed operating conditions of these computational experiments are listed in Table 1. To account for the high measurement noise and irregular sample pattern expected for industrial experiments, data points were collected at irregular time steps and a 10% error (denoting measurement error in real plants) was applied. Raw data generation was implemented in Mathematica 11.

$$\frac{dc_X}{dt} = u_0 \cdot \frac{c_N}{c_N + K_N} \cdot c_X - u_d \cdot c_X \tag{1a}$$

$$\frac{dc_N}{dt} = -Y_{N/X} \cdot u_0 \cdot \frac{c_N}{c_N + K_N} \cdot c_X + F_{in} \cdot c_{N,in} \tag{1b}$$

$$\frac{dc_L}{dt} = k_0 \cdot \frac{c_N}{c_N + K_{NL}} \cdot c_X - k_d \cdot c_L \cdot c_X \tag{1c}$$

$$I(z) = I_0 \cdot \left(e^{-(\tau \cdot X + K_a) \cdot z} + e^{-(\tau \cdot X + K_a) \cdot (L - z)}\right) \tag{1d}$$

$$u_0 = \frac{u_m}{20} \cdot \sum_{n=1}^{9} \left(\frac{I_0}{I_0 + k_s + \frac{I_0^2}{k_i}} + 2 \cdot \frac{I_{\frac{n \cdot L}{10}}}{I_{\frac{n \cdot L}{10}} + k_s + \frac{I_{\frac{n \cdot L}{10}}^2}{k_i}} + \frac{I_L}{I_L + k_s + \frac{I_L^2}{k_i}} \right) \tag{1e}$$

$$k_0 = \frac{k_m}{20} \cdot \sum_{n=1}^{9} \left(\frac{I_0}{I_0 + k_{sL} + \frac{I_0^2}{k_{iL}}} + 2 \cdot \frac{I_{\frac{n \cdot L}{10}}}{I_{\frac{n \cdot L}{10}} + k_{sL} + \frac{I_{\frac{n \cdot L}{10}}^2}{k_{iL}}} + \frac{I_L}{I_L + k_{sL} + \frac{I_L^2}{k_{iL}}} \right) \tag{1f}$$

where c_X, c_N, and c_L are concentration of biomass, nitrate, and lutein, respectively, F_{in} and $c_{N,in}$ are nitrate inflow rate and concentration, respectively, z is distance from light source, L is width of the reactor. Other kinetic parameters are listed in Table 2.

2.3.2 Generation of high quality data

The raw datasets obtained in Section 2.3.1 contain high levels of noise and are incomplete (*i.e.* missing data points). To enable their use for the construction of the data-driven model, they are initially sent to the middle-level of the framework to filter out noise and add missing data points by using a simple kinetic model. The simple model selected in this work is a hybrid of the Monod and Logistic models (Del Rio‐Chanona et al., 2018). Its structure is presented in Eq. (2a)-(2c), with symbols denoting same physical meanings to those in Eq. (1a)-(1f). Compared to the original mechanistic model which has 15 parameters, this simple kinetic model only contains 6 parameters with a much less nonlinear structure. Given that the data-fitting capability of kinetic models is limited to specific sets of operating conditions, *it is essential to process each dataset individually*. This means that the raw datasets should be fed into the kinetic model *separately*, so that the best set of model parameters is identified for each single dataset. This allows the kinetic model to retrieve as much of the authentic behaviours of each process as possible, maximising the quality of the corrected datasets.

$$\frac{dc_X}{dt} = u_0 \cdot \frac{c_N}{c_N + K_N} \cdot c_X - u_d \cdot c_X^2 \quad (2a)$$

$$\frac{dc_N}{dt} = -Y_{N/X} \cdot \left(u_0 \cdot \frac{c_N}{c_N + K_N} \cdot c_X - u_d \cdot c_X^2 \right) + F_{in} \cdot c_{N,in} \quad (2b)$$

$$\frac{dc_L}{dt} = Y_{L/X} \cdot u_0 \cdot \frac{c_N}{c_N + K_N} \cdot c_X - k_d \cdot c_L \cdot c_X \quad (2c)$$

where $Y_{L/X}$ is lutein yield coefficient and μ_0 is average specific cell growth rate.

In practice, data rectification was addressed by successively estimating the kinetic model parameters for each raw dataset using a nonlinear least-squares optimisation algorithm. The model was discretised and transformed into a nonlinear programming problem. Orthogonal collocation over finite elements in time was used as the discretisation method. Values of parameters were estimated using the interior point nonlinear optimisation solver IPOPT. The execution was programmed in Python using *Pyomo* (Hart, Laird, Watson, & Woodruff, 2012). As the model was used to fit each dataset separately, in total, parameter estimation was carried out *three* times to generate *three* sets of parameter values. The results are shown in Section 3.

2.3.3 Data-driven model construction

The successful construction of data-driven models is strongly dependent on how the training data is formed to seek the best model structure. Hence, once high-quality datasets (replacing the original three sets) were generated at the framework middle-level, they were utilised to build a high-fidelity machine learning based model. The high-quality datasets have regular time steps to facilitate the construction of the data-driven model, set to 6 hours in the present work. Artificial neural network (ANN) was chosen as the data-driven model. It was designed to predict changes of state variables within one time interval from the current time, using the previous values of state variables and nitrate feed flowrates. By successively applying this ANN, future values of state variables are predicted multiple time steps ahead.

ANN requires a large number of datasets to guarantee its interpolation power (Baughman & Liu, 1995), and its accuracy greatly depends on the selection of model hyperparameters (*e.g.* number of neurons, layers, training epochs). Thus, several strategies developed in our recent research have been adopted here. The first is to create a large size of artificial datasets by embedding adequate random noise into the three sets of high quality data. *It is worth noting that this random noise is used for ANN training (Level 3) and should not be confused with the measurement error present in the low quality data obtained from the real plant (Level 1)*. The augmentation of datasets has been confirmed to significantly reinforce ANN's accuracy and relief the pressure from real plant data acquisition (Tulsyan et al., 2018). Data augmentation can also be interpreted as a regularisation strategy, which is far more intuitive, and hence easier to implement correctly, than traditional penalising strategies. The magnitude of random noise is specific to each system; it should be large enough for the ANN to distinguish between each dataset but small enough not to disguise the real process behaviour. In this study, this random noise was identified to be 5%, and 100 artificial datasets were generated for each high quality dataset. Once augmented, all datasets were normalised before employing them to train the ANN. An alternative strategy would be the use of advanced artificial data augmentation techniques *e.g.* Generative Adversarial Networks (GANs) (Antoniou, *et. al*, 2017). However, as they involve more complicated training frameworks and their applicability to chemical processes has not been investigated, they were not used here but will be explored in the future.

The second strategy is to adopt a hyperparameter selection framework which determines the optimal trade-off between *variance* and *bias*. Increasing hyperparameters complicate the model structure, leading to a better fit of training datasets. However, this will increase the likelihood of over-fitting and worsen the model's prediction accuracy. In addition, higher model complexity also causes higher computational cost. Another strategy used is the *k*-fold

method, where a selection of k-1 from the k datasets is used to train an ANN and the remaining dataset is used to evaluate the maximum prediction error of this model. Then, another k-1 set is selected to repeat this procedure until the best model is found. Detailed explanations of these methods and their implementations have been illustrated in recent studies (Bradford et al., 2018; Del Rio‐Chanona et al., 2018). Data augmentation and ANN construction were executed in Mathematica 11.

2.4 Online optimisation using the hybrid modelling framework

To demonstrate the performance of this hybrid modelling framework, a new computational experiment was designed for an online optimisation framework. Consistent with the initial dataset generation experiments (Section 2.3.1), this experiment was run for a total duration of 7 days and a nitrate feed was established after 48 hours. The hybrid modelling framework was used to administer the optimal actions (nitrate inflow rate). Similar to real industrial implementations, nitrate inflow rate in this experiment can change once every 6 hours. The detailed operating conditions and bounds on nitrate inflow rates are shown in Table 1. For the optimisation of the process, this framework was incorporated into an EMPC or a MPC framework. A brief overview of EMPC and MPC is presented next.

2.4.1 Introduction to model predictive control strategies

Model predictive control (MPC) has been widely used in chemical industries. It contains four levels as shown in Fig. 2(a). The top-level determines set-points by means of steady-state optimisation from the plant overarching decisions. These high-level set-points are used to find a temporary trajectory (process set-points) over a certain horizon at a unit level. During the ongoing process, future optimal control actions within this horizon are predicted via a dynamic model in the MPC-level to follow the set-points and are delivered to the bottom-level for execution at the next time step. This procedure is repeated successively to renew control decisions (Maciejowski, 2002). MPC is applied in steady-state operation and requires

the availability of a process reference trajectory. This, however, cannot be easily satisfied in bioprocesses as they are often operated dynamical. Due to the complexity and stochastic nature of metabolic networks, it is unlikely that MPCs can predict optimal trajectories which can be easily followed by the bioprocess. Hence, economic model predictive control (EMPC) was developed to replace MPC by applying an economic index to merge the set-point optimisation and predictive control levels (Ehecatl Antonio del Rio-Chanona, Zhang, et al., 2016), reducing the number of framework levels to three (Fig. 2(a)).

When applying model predictive control strategies for the online optimisation of bio-production systems, it is particularly important that the dynamic model is accurate. Therefore, a finite-data estimation window least-squares method (FDWLS) was integrated into the EMPC framework in this work so that newly measured plant data can be used to improve the model accuracy (Fig. 2(b)). After each time step, and prior to identifying future control actions, data acquired from the ongoing process (*i.e.* time *T-k* to *T* in Fig. 2(b)) was used to tune the data-driven model and minimise model-plant mismatch. The modified model was then utilised to predict optimal operating conditions over the next control horizon. Through this integrated framework, model accuracy and optimal control actions can be frequently synchronised. Finally, the top-level, plant-wide static optimisation, in MPC and EMPC was excluded as the current study does not cover system-wide online optimisation.

2.4.2 Identification of future control actions using the data-driven model

The ANN was used to predict optimal control actions during the ongoing process. For EMPC, the objective is to maximise lutein production at the end of the control horizon. A penalty was also added to punish large variations between adjacent control actions and to alleviate the sensitive response of algal metabolic network. The EMPC problem is shown in Eq. (3a)-(3b).

$$\max_{F_{in}(t)} \quad c_L(t_{FMPC}) + \sum_{i=1}^{N-1} \left(\Delta F_{in,i}(t)\right)^2 \cdot (-\sigma_i) \tag{3a}$$

subject to:

process dynamics (formulated by the ANN)

$$0.0 \leq F_{in}(t) \leq 1000 \text{ µL h}^{-1} \tag{3b}$$

where t_{FMPC} is the final time of the control horizon, $\Delta F_{in,i} = F_{in,i+1} - F_{in,i}$, $F_{in,i}$ is the i^{th} control action in the control horizon, N is the total number of control actions in the control horizon, $\sigma_i > 0$ is the penalty coefficient to reduce deviation between the two adjacent controls, and $c_{L,R}(t_{FMPC})$ is lutein production in Exp. 3.

To test the effectiveness of this modelling framework for MPC, the track of a previous computational experiment that yields the highest lutein production (Exp. 3) was used as the MPC reference trajectory. Given that the new process aims to enhance lutein production instead of reproducing previous results, the MPC objective was set such that at each time step lutein production in the ongoing process at the end of the control horizon (t_{EMPC}) should be close to 1.2 times of that in Exp. 3 (based on the maximum theoretical production predicted by offline optimisation in Section 3.2.2). This objective function is formulated as Eq. (3c).

$$\min_{F_{in}(t)} \quad \left(c_L(t_{FMPC}) - 1.2 \cdot c_{L,R}(t_{FMPC})\right)^2 + \sum_{i=1}^{N-1} \left(\Delta F_{in,i}(t)\right)^2 \cdot \sigma_i \tag{3c}$$

The length of the control (and prediction) horizon in this work was fixed to 2 days, similar to the industrial setup, thus the framework estimated 8 control actions during each iteration. A hybrid stochastic optimisation algorithm was designed to optimise the ANN chain, where random search was initially executed to narrow down the solution space and then simulated annealing was applied to refine the optimal solution (Ehecatl Antonio del Rio-Chanona et al., 2018). This was executed in Mathematica 11. *Discrete* future process behaviours (*i.e.* future data points) were also simultaneously estimated by the ANN once control actions were determined.

2.4.3 Visualisation of future process trajectory

The simple kinetic model (Eq. (2a)-(2c)) in the middle-level of the hybrid modelling framework was then used to generate the *continuous* future process behaviours. This was achieved by re-fitting its model parameters using the ANN predicted data points. In such a way, full future trajectories of state variables are visualised throughout the control horizon, and the simple kinetic model acts as a soft sensor. Through this approach, samples from the ongoing plant can be taken at any time for monitoring purposes, improving the flexibility of data measurement and analysis. Measured data were then compared against predicted process trajectories to verify the accuracy of the hybrid modelling framework. Finally, new data measured from the plant were collected at the end of each day to refine the accuracy of the modelling framework through FDWLS, and optimal control actions over the next 24 hours were updated and executed into the ongoing process. The length of the FDWLS was chosen as 2 days. Process trajectory visualisation and FDWLS were carried out in Mathematica 11.

3. Results and discussion

3.1 Results of the hybrid modelling framework construction

3.1.1 Results of high quality data generation

The three sets of parameter values for the simple kinetic model fitted to the three sets of raw computational data are listed in Table 3, and fitting results for individual datasets are shown in Fig. 3. From Table 3, it is seen that values of most model parameters change substantially between the three different datasets (except for u_0 and k_d), confirming the challenge of using a single set of parameter values to simulate the distinct behaviours of same system operated under different conditions. The excellent data fits displayed in Fig. 3 show that the simple kinetic model does not only fit all three individual sets of experimental data (Fig. 3(c)), but that it is also capable of effectively removing large levels of noise as the model regression curves closely follow the "true" data points (without measurement error) (Fig. 3(a) and 3(b)). In specific, average deviations between the model simulation result and the true data for

concentrations of biomass, lutein, and nitrate are 4.6%, 4.1%, and 8.4%, respectively, with the largest deviation being 5.8%, 5.2%, and 12.6%, respectively. Thus, it can be concluded that a simple kinetic model can well screen raw datasets for high quality data generation.

3.1.2 Results of the data-driven model construction

As described in Section 2.3.3, high quality datasets were extracted from the regression results of the simple kinetic model at fixed time intervals of 6 hours. Through the use of different hyperparameter selection strategies, the best structure of the current ANN was found to enclose two hidden layers with 4 neurons in the input layer, 3 neurons in the output layer, 20 neurons in the first hidden layer, and 15 neurons in the second hidden layer. As the ANN has only been trained with changes of state variables over one time interval (6 hours), it is essential to verify its predictive capability over a longer time span (*i.e.* 48 hours). This is because the purpose of the hybrid modelling framework for online optimisation is to identify optimal control actions over the entire control horizon (2 days).

Hence, the ANN was tested by predicting the process behaviours of the three computational experiments (Exp. 1 to Exp. 3) over the whole operation time course using *only* the initial conditions and nitrate feed rates (Fig. 4). It can be seen that the ANN can accurately predict the entire process behaviour under different operating conditions, indicating its great predictive power for process online optimisation. Attention should be paid to the fact that ANN is not the only data-driven model that can be used in this framework. If the process is governed by multiple physical mechanisms, *e.g.* biological kinetics and fluid dynamics, other models *e.g.* recurrent neural network can also be applied (Baughman & Liu, 1995).

3.2 Results of online optimisation

3.2.1 Accuracy of the modelling framework for online optimisation

The modelling framework was used to optimise lutein production in the new computational fed-batch process (Online Exp. in Table 1). Consistent with industrial operational practices,

the process was initially operated as a batch system to facilitate biomass growth. After 48 hours a nitrate feed was established to replenish the concentration of this nutrient inside the reactor to maintain cell growth and lutein production until Day 7 (feed rate re-adjusted every 6 hours). The prediction results of the current framework integrated with FDWLS and EMPC after 48 hour are shown in Fig. 5.

At the end of Day 2, the modelling framework was retuned using the process data obtained over the previous day using FDWLS, before the ANN was used to predict the optimal nitrate feed rates and the associated discrete process behaviours for Days 3 and 4 (black points in Fig. 5). These data points were then fed into the simple kinetic model to generate continuous process trajectories over these days (lines in Fig. 5) and the new estimated values of the kinetic model parameters (listed in Table 3). Consistent with the conclusion from Section 3.1.1, the values of these parameters are significantly different compared to those from the previous experiments. On Day 3, the optimal control actions predicted after 48 hours (end of Day 2) were implemented into the ongoing process. Over the course of this day, the process samples were measured by the ongoing computational experiment at irregular time intervals (red points in Fig. 5) to compare against the framework predictions.

From the figure, it is seen that these measurements (containing 10% error) are closely aligned to the framework predictions, suggesting the high accuracy and predictive capability of the hybrid modelling framework. In addition, average deviations between the visualisation result of the re-fitted simple kinetic model and the "true" process behaviour (calculated using the complex mechanistic model without 10% measurement error) are 5.1%, 11.7%, and 2.6% for the concentration of biomass, nitrate, and lutein, respectively. The modelling framework alongside the FDWLS and EMPC frameworks was then repeated until the end of the process.

3.2.2 Efficiency of different process optimisation strategies

To analyse the performance of EMPC and MPC in the current system, the original complex model (Eq. (1a)-(1f)) which was used to predict the computational datasets used in this study was also utilised to optimise the new computational process *prior to* its experimentation. This optimisation approach is known as offline optimisation (open-loop optimisation) and predicts the best theoretically possible optimisation scheme (without any process disturbances or model-plant mismatches). Hence, the offline optimisation result was used as the benchmark for examining the online optimisation performance. In addition, to explore the maximum lutein production, the penalty term in Eq. (3a), which accounts for the dramatic changes of control actions, was removed from the offline optimisation. However, it is important to stress that in most cases offline optimisation is not feasible due to a lack of accurate complex kinetic models and the frequent occurrence of process disturbances.

The optimisation results of EMPC and the open-loop optimisation are compared in Fig. 6. It is seen that both schemes give similar final lutein yields of 5.0 mg L^{-1} (Fig. 6(c)). However, for most of the production period, the process optimised by EMPC shows the higher lutein yields except for the final stages of operation. This is explained by the design of the EMPC, which aims to optimise lutein production at the end of the control horizon (48 hours) instead of the end of the entire operation. Although extending the duration of the control horizon may improve the overall control scheme, it will increase the computational requirements of the process optimiser making it more expensive and difficult to manage. As a result, this trade-off should be carefully balanced. It is also observed that control action changes determined by the EMPC framework are less abrupt than those from the offline optimisation scheme (Fig. 6(d)). This is particularly advantageous for the optimisation of bioprocesses, where rapid changes to the culture environment can greatly disturb the microbial metabolic activities causing the system to become unstable. Thus, the high lutein production predicted through the open-loop optimisation may not be achievable in practice.

The MPC guided process was found to give similar lutein yields to the EMPC derived process, thus not presented in detail. Practical concerns associated with the two optimisation approaches (*e.g.* stability) have been described in previous work (Maciejowski, 2002). It is, however, important to stress that it is challenging to find a suitable MPC reference trajectory for general bioprocesses. One practical reason is that it is difficult to guarantee consistent initial operating conditions and initial cellular metabolic activities between different batches, both of which have decisive impacts on cell growth and product formation. As a result, EMPC may represent the better choice for industrial bioprocess online optimisation.

Conclusion

Overall, it is concluded that this hybrid modelling framework represents an effective strategy to resolve practical issues associated with the optimisation of industrial biosystems, including the low quality and quantity of available data, lack of knowledge of the physical mechanisms of the process, high costs of frequent sampling and online measurements, and challenges in pre-determining set-points for fed-batch operations. When combined with advanced model adaptation and online optimisation schemes, this study has shown that the hybrid modelling framework provides high predictive and flexible capabilities, indicating its suitability for industrial application. Most importantly, the optimal results from the online optimisation framework are almost identical to those obtained from the best theoretically possible optimisation scenario which does not account for system disturbances, measurement noise, or model-plant mismatches. This directly suggests the potential of the current framework for the simulation and optimisation of complex biosystems. Moreover, this framework can be easily adapted to include multi-objective optimisation to find optimal solutions based on different criteria, such as operating conditions and other system requirements.

Regarding the structure of the hybrid modelling framework, it is important to note that the simple kinetic model provides a critical function to enable the use of low quality raw data for

the construction of the high-fidelity data-driven model as well as the visualisation of future process behaviours during online optimisation. It is envisaged that through the development of cutting-edge hybrid modelling techniques, the design of efficient optimisation algorithms for data-driven models will become a critical task for future biochemical engineering research. Moreover, it is worth highlighting that the current hybrid modelling framework aims to provide a general structure for industrial bioprocess digitalisation and optimisation, and it can take advantages of different physical and data-driven modelling techniques to improve its capability. For instance, autoassociative neural network (AANN) can be used to replace the simple kinetic model for noise filtering (Baughman & Liu, 1995); current advances in machine learning based dynamic model structure discovery can be implemented at the top-level to identify the best physical model structure for process prediction and visualisation; reinforcement learning can be used as an alternative approach for automatic process optimal control (Petsagkourakis, Sandoval, Bradford, Zhang, & del Rio-Chanona, 2019). Other techniques such as Gaussian processes (Bradford et al., 2018; Tulsyan et al., 2018) can be also adopted to estimate the uncertainty of model predictions for product quality control and process monitoring. The best combination of these modelling, visualisation and optimisation strategies should be extensively studied to consolidate efficiency of the current hybrid modelling framework.

Acknowledgement

This project has received funding from the EPSRC project (EP/P016650/1).

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

Antoniou, A., Storkey, A., & Edwards, H. (2017). Data Augmentation Generative Adversarial Networks. Retrieved from http://arxiv.org/abs/1711.04340

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

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

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

del Rio-Chanona, E. A., Ahmed, N. rashid, Zhang, D., Lu, Y., & Jing, K. (2017). Kinetic modeling and process analysis for Desmodesmus sp. lutein photo-production. *AIChE Journal*, *63*(7), 2546–2554. https://doi.org/10.1002/aic.15667

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

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

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

del Rio-Chanona, E. A., Zhang, D., & Vassiliadis, V. S. (2016). Model-based real-time optimisation of a fed-batch cyanobacterial hydrogen production process using economic model predictive control strategy. *Chemical Engineering Science*, *142*, 289–298. https://doi.org/10.1016/j.ces.2015.11.043

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

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

Fouchard, S., Pruvost, J., Degrenne, B., Titica, M., & Legrand, J. (2009). 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. *Biotechnology and Bioengineering*, *102*(1), 232–277. https://doi.org/10.1002/bit.22034

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

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

Jeandet, P., Vasserot, Y., Chastang, T., & Courot, E. (2013). Engineering Microbial Cells for the Biosynthesis of Natural Compounds of Pharmaceutical Significance. *BioMed*

Research International, *2013*, 1–13. https://doi.org/10.1155/2013/780145

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

Maciejowski, J. M. (2002). *Predictive Control: With Constraints*. Prentice Hall.

Marchetti, A., François, G., Faulwasser, T., & Bonvin, D. (2016). Modifier Adaptation for Real-Time Optimization—Methods and Applications. *Processes*, *4*(4), 55. https://doi.org/10.3390/pr4040055

Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D., & del Rio-Chanona, E. A. (2019). Reinforcement Learning for Batch Bioprocess Optimization. Retrieved from http://arxiv.org/abs/1904.07292

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

Valdez-Castro, L., Baruch, I., & Barrera-Cortés, J. (2003). Neural networks applied to the prediction of fed-batch fermentation kinetics of Bacillus thuringiensis. *Bioprocess and Biosystems Engineering*, *25*(4), 229–233. https://doi.org/10.1007/s00449-002-0296-7

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

Voll, A., & Marquardt, W. (2012). Reaction network flux analysis: Optimization-based evaluation of reaction pathways for biorenewables processing. *AIChE Journal*, *58*(6), 1788–1801. https://doi.org/10.1002/aic.12704

Wagner, J. L., Lee-Lane, D., Monaghan, M., Sharifzadeh, M., & Hellgardt, K. (2019).

Recovery of excreted n-butanol from genetically engineered cyanobacteria cultures: Process modelling to quantify energy and economic costs of different separation technologies. *Algal Research*, *37*, 92–102. https://doi.org/10.1016/j.algal.2018.11.008

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

Zhang, D., Dechatiwongse, P., del Rio-Chanona, E. A., Maitland, G. C., Hellgardt, K., & Vassiliadis, V. S. (2015). Modelling of light and temperature influences on cyanobacterial growth and biohydrogen production. *Algal Research*, *9*, 263–274. https://doi.org/10.1016/j.algal.2015.03.015

Zhang, D., del Rio-Chanona, E. A., & Shah, N. (2017). Screening Synthesis Pathways for Biomass-Derived Sustainable Polymer Production. *ACS Sustainable Chemistry & Engineering*, *5*(5), 4388–4398. https://doi.org/10.1021/acssuschemeng.7b00429

Zhang, D., Wan, M., del Rio-Chanona, E. A., Huang, J., Wang, W., Li, Y., & Vassiliadis, V. S. (2016). Dynamic modelling of Haematococcus pluvialis photoinduction for astaxanthin production in both attached and suspended photobioreactors. *Algal Research*, *13*(12), 69–78. https://doi.org/10.1016/j.algal.2015.11.019

Table 1: Operating conditions of the computational experiments (Exp. 1-3, used for data generation) and the online computational experiment.

	Exp. 1	Exp. 2	Exp. 3	Online Exp.
Initial biomass conc. (g L^{-1})	0.2	0.2	0.2	0.2
Incident light intensity (μmol m^{-2} s^{-1})	750	750	750	750
Nitrate feed conc. (mol L^{-1})	0.2	0.2	0.2	0.2
Nitrate feed rate (μL h^{-1})	100	500	1000	[0.0, 1000]
Operating time (h)	168	168	168	168

Table 2: List of kinetic model parameters. Their values can be found in the previous work (del Rio-Chanona *et al.*, 2017).

u_0	cell specific growth rate	u_d	cell specific decay rate
$Y_{N/X}$	nitrate yield coefficient	k_0	lutein synthesis rate constant
k_d	lutein consumption rate constant	I_0	incident light intensity
k_s	light saturation term (cell growth)	k_{sL}	light saturation term (lutein synthesis)
k_i	light inhibition term (cell growth)	k_{iL}	light inhibition term (lutein synthesis)
τ	cell absorption coefficient	K_a	bubble scattering coefficient
K_N, K_{NL}	nitrate half-velocity constant for cell growth and lutein synthesis, respectively		
u_m, k_m	maximum specific growth rate and lutein synthesis rate constant, respectively		
I_i	local light intensity at a distance of $\frac{n \cdot L}{10}$ from the bioreactor exposure surface		

Table 3: Values of kinetic model parameters estimated using different experimental dataset

Parameter	Exp. 1	Exp. 2	Exp. 3	Online experiment
u_0, h^{-1}	0.0268	0.0243	0.0231	0.0223
u_d, L g^{-1} h^{-1}	0.0267	8.85×10^{-3}	7.45×10^{-3}	3.30×10^{-3}
$Y_{N/X}$, mg g^{-1}	560.4	444.1	571.4	505.1
K_N, mg L^{-1}	37.13	3.978	0.977	6.416
$Y_{L/X}$, mg g^{-1}	4.420	4.048	4.112	5.698
k_d, L g^{-1} h^{-1}	0.021	0.018	0.017	0.024

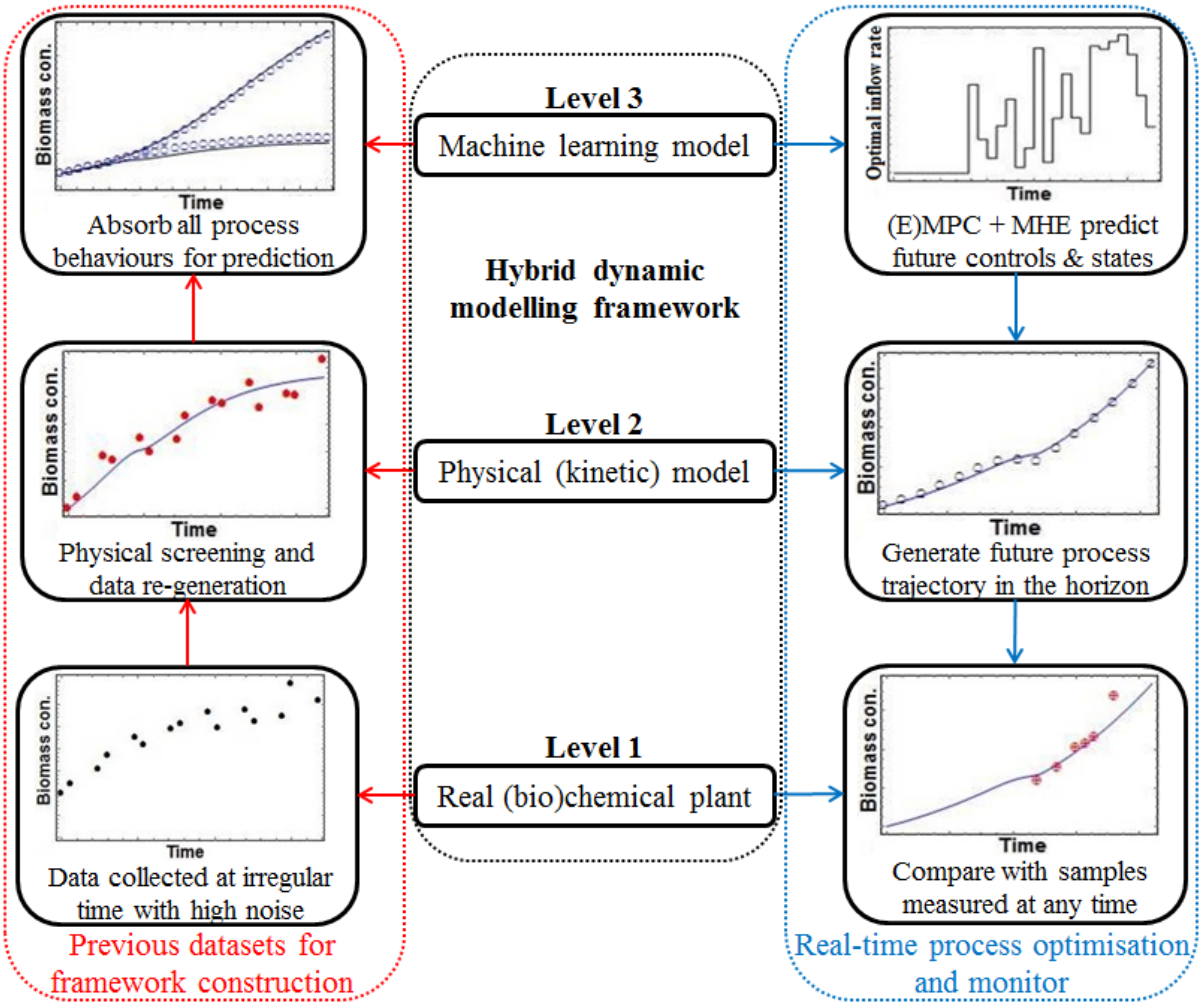

Figure 1: Schematic of the hybrid modelling framework.

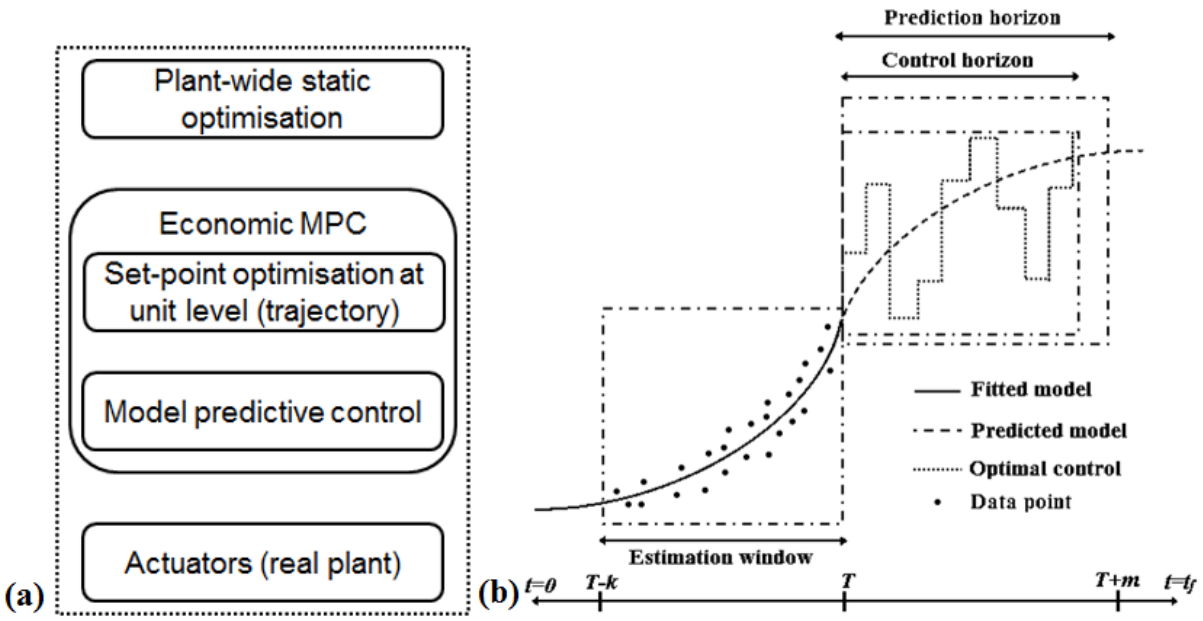

Figure 2: Schematics of MPC, EMPC, and FDWLS. (a): A general framework of MPC and EMPC. (b): The integrated framework of FDWLS and EMPC.

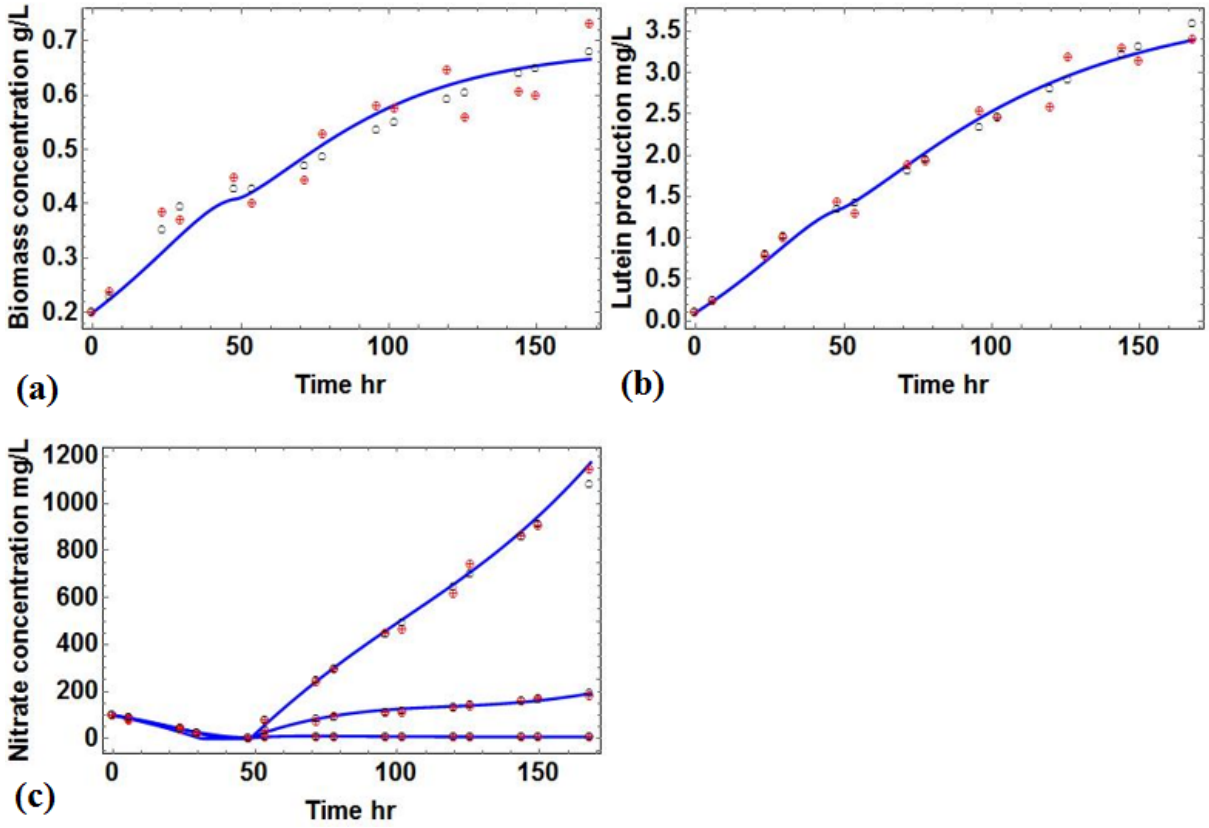

Figure 3: Fitting result of the simple kinetic model. Line: kinetic model simulation (regression) result. Red point: raw data points (with 10% measurement error). Black point (open circle): "true" data points without measurement error. (a) and (b): Kinetic model fitting result of biomass concentration and lutein production for Exp. 1, respectively. (c): Nitrate concentration fitted by the kinetic model for all the three experiments (Exp. 1-Exp. 3).

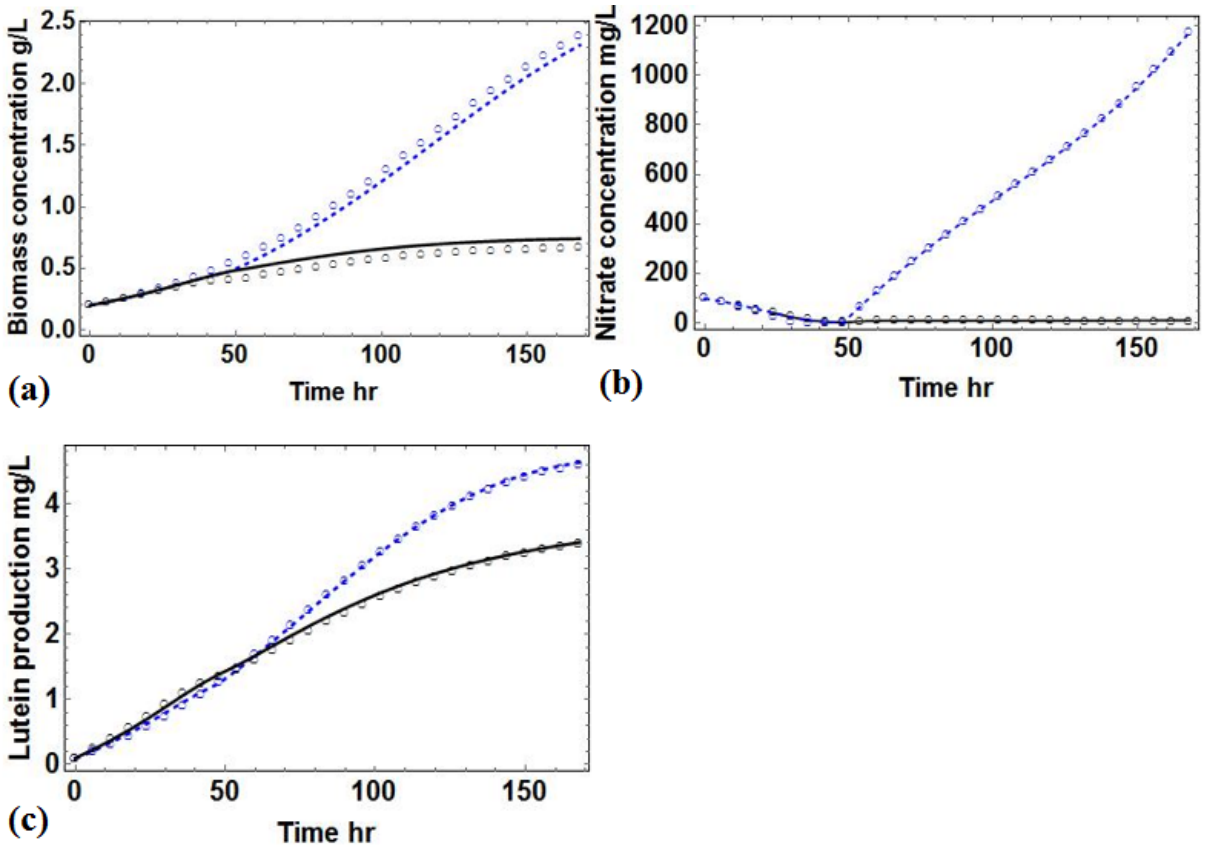

Figure 4: ANN simulation result over the entire experimental operation course using only initial operating conditions and nitrate inflow rates. (a): biomass concentration, (b): nitrate concentration, (c): lutein production. Points: ANN prediction result of Exp. 1 (black) and Exp. 3 (blue). Lines: Continuous process behaviour of Exp. 1 (black) and Exp. 3 (blue) simulated based on the simple kinetic model (high quality data).

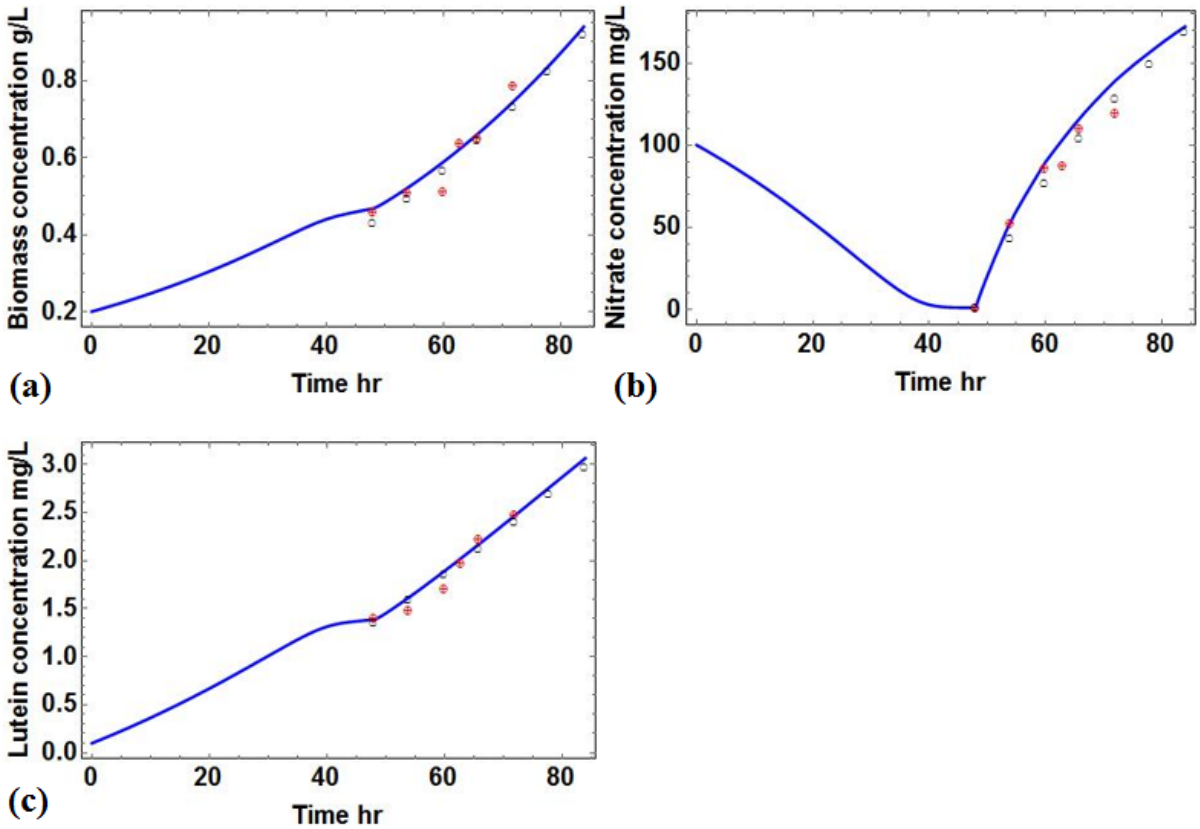

Figure 5: Prediction of the hybrid modelling framework after 48 hours (end of Day 2). (a): biomass concentration. (b): nitrate concentration. (c): lutein production. Black (empty) points: data-driven model prediction (*discrete* future process behaviours) over a fixed time interval (48^{th} – 84^{th} hour). Line: *continuous* future process behaviours visualised by the simple kinetic model (soft sensor) through parameter re-estimation. Red points: samples measured from the plant at random times during Day 3 (48^{th} – 72^{nd} hour) with 10% measurement error.

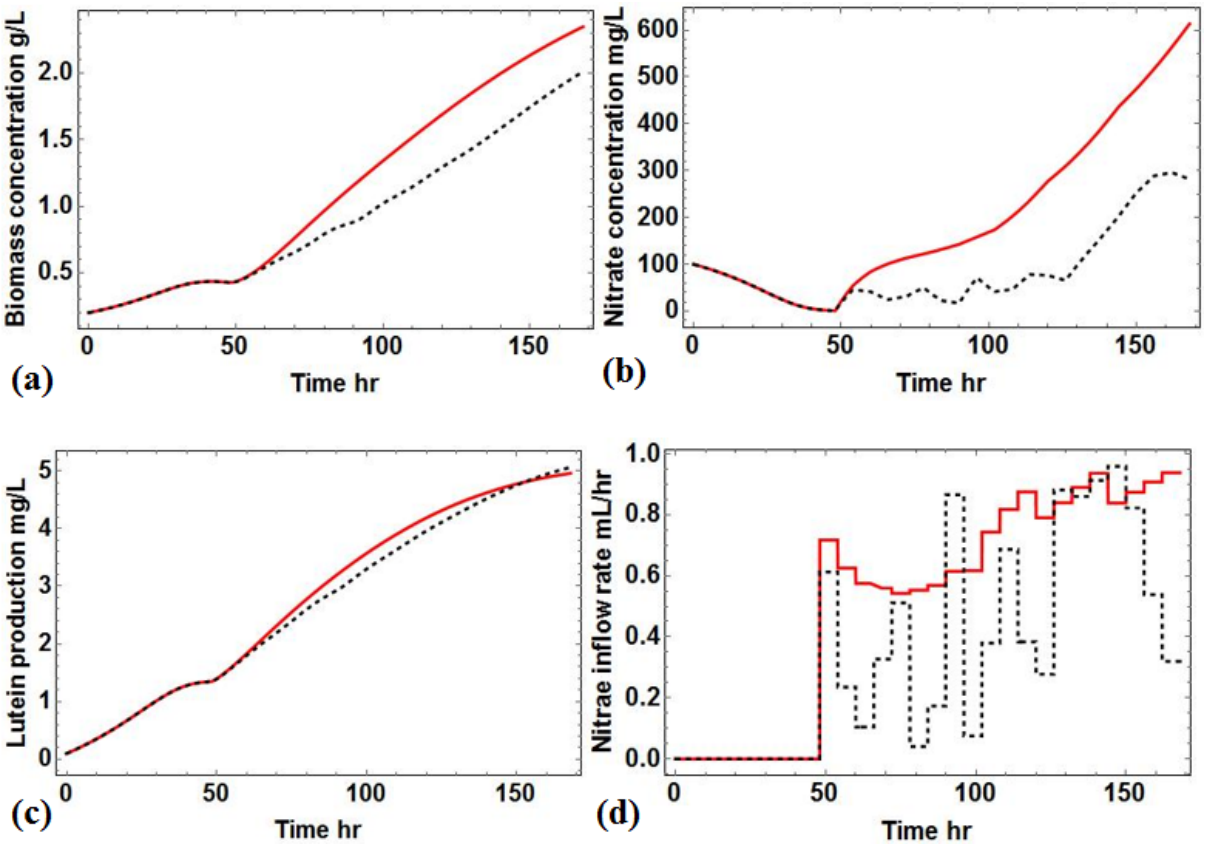

Figure 6: Optimisation result of the EMPC guided process and the open-loop optimisation derived process (the theoretically best process performance without process disturbance, measurement noise, and model-process mismatch). (a): biomass concentration. (b): nitrate concentration. (c): lutein production. (d): nitrate inflow rate. Solid line: trajectory of the EMPC derived process; dashed line: trajectory of the open-loop optimisation derived process.