AIChE Journal
DOI 10.1002/aic.16473

Deep Learning based Surrogate Modeling and Optimization for Microalgal Biofuel Production and Photobioreactor Design

Ehecatl Antonio del Rio-Chanona[1,2‡], Jonathan L. Wagner[2,3‡], Haider Ali[4], Fabio Fiorelli[5], Dongda Zhang[1,2,6,7*], Klaus Hellgardt[2]

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

2: Dept. of Chemical Engineering, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK

3. Dept. of Chemical Engineering, University of Loughborough, Loughborough, LE11 3TU, UK

4: School of Mechanical Engineering, Kyungpook National University, 1370 Sankyuk-Dong, Buk-gu Daegu 702701, South Korea

5: Materialize X Ltd., 1 Fyfield Road, London, SW9 7HW, UK

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

7: School of Chemical Engineering and Analytical Science, University of Manchester, Oxford Road, Manchester, M1 3AL, UK

‡: These authors contributed equally to this work.

*: corresponding author, email: dongda.zhang@manchester.ac.uk.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/aic.16473

Received: Apr 11, 2018; Revised: Oct 08, 2018; Accepted: Nov 09, 2018

Abstract

Identifying optimal photobioreactor configurations and process operating conditions is critical to industrialize microalgae-derived biorenewables. Traditionally, this was addressed by testing numerous design scenarios from integrated physical models coupling computational fluid dynamics and kinetic modelling. However, this approach presents computational intractability and numerical instabilities when simulating large-scale systems, causing time-intensive computing efforts and infeasibility in mathematical optimization. Therefore, we propose an innovative data-driven surrogate modelling framework which considerably reduces computing time from months to days by exploiting state-of-the-art deep learning technology. The framework built upon a few simulated results from the physical model to learn the sophisticated hydrodynamic and biochemical kinetic mechanisms; then adopts a hybrid stochastic optimization algorithm to explore untested processes and find optimal solutions. Through verification, this framework was demonstrated to have comparable accuracy to the physical model. Moreover, multi-objective optimization was incorporated to generate a Pareto-frontier for decision-making, advancing its applications in complex biosystems modelling and optimization.

Keywords: surrogate modelling, convolutional neural network, hybrid stochastic optimization, excreted biofuel, photobioreactor design

Introduction

For a number of decades, microalgae have attracted significant interest as feedstocks for the production of renewable bioenergy and sustainable high-value bioproducts [1,2]. One of their main advantages is their ability to directly utilize solar energy to convert atmospheric CO_2 into biorenewable products, ranging from biofuels (*e.g.* biodiesel, biohydrogen) [3,4] to valuable food additives and pharmaceutical ingredients (*e.g.* astaxanthin, lutein) [5,6]. Moreover, recent development of genetically engineered algal strains, which are able to excrete bioproducts directly into the medium, show great potential to further reduce the downstream biorenewables' separation cost [7,8].

To facilitate the industrialisation of microalgal bioprocesses, extensive research has been conducted on the identification of optimal operating conditions including light intensity, temperature, and nutrient supply, in order to achieve maximum biomass growth and bioproduct synthesis [9–11]. A major challenge to commercialization is light attenuation (decrease of local light intensity along the light transmission direction) caused by algal cell absorption and bubble scattering, critically limiting the potential for high density biomass cultivation [12]. Consequently, a range of different types of photobioreactors (PBRs) have been designed to improve the light distribution for large scale operation [13].

Meanwhile, the rapid development of computer-aided technology has resulted in extensive simulation research on microalgae bioprocess modelling and optimization. For instance, both kinetic models and machine learning based models have been proposed for the simulation of biorenewables' production [14,15]. In addition, computational fluid dynamics (CFD) have been implemented to assist the design of low cost and high efficiency PBRs [16]. These studies indicated that microalgal biomass cultivation is predominantly affected by two fluid dynamic

factors; shear stress and liquid mixing [17,18]. While rapid mixing of liquid cultures along the direction of light transmission can improve the overall light utilization, thereby enhancing photosynthetic efficiencies [19], it also induces intense shear stresses which can severely damage algal cells resulting in biomass death [18,20].

Therefore, in order to transition microalgae based bioprocesses from laboratory to industrial scale, it is essential to identify suitable PBR configurations together with optimized operating conditions to enable large scale microalgae cultivation and biorenewable production. Due to the high costs associated with experimental investigations, it is commonly accepted that accurate simulations are required to support this transition. However, the complexity of the interactive hydrodynamics and kinetic mechanisms makes it challenging to construct a multi-scale physical model which is capable of simulating cell growth and product synthesis accurately in large scale PBRs. Although a few studies have attempted to couple kinetic models and CFD models [21,22], these have been restricted to the analysis of open ponds instead of PBRs. Moreover, the execution of mathematical optimization on multi-scale physical models is infeasible in most cases due to the high nonlinearity and stiffness of the models; while using the integrated models for the sampling of large numbers of different combinations of reactor configurations and operating conditions is also impractical due to the substantial computational costs of running each design scenario. As a result, little improvement has been achieved in this domain to date.

The current study proposes an efficient surrogate modelling framework to resolve the aforementioned challenges and simultaneously optimize the operating conditions and PBR configuration of a pilot scale microalgal biofuel production unit. In particular, this framework consists of a convolutional neuron network (CNN)-based surrogate model that can learn from

an existing physical model to predict unknown process behaviours accurately. This modelling strategy is coupled with a hybrid stochastic optimization algorithm that can efficiently explore optimal solutions using the current data-driven model. The advantage of this framework is that it can reduce substantially the computing time required for optimal solution estimation (*e.g.* from months to days) compared to the conventional physical modelling approach and meanwhile guarantees the high quality of its identified solution. In the following sections the detailed modelling framework procedure and optimization results are presented together with their associated advantages.

Methodology

Introduction of modelling framework

The modelling framework proposed in the current study includes 6 steps:

1. Construct a CFD model and a kinetic model to simulate PBR hydrodynamics and bioprocess kinetic behaviours, respectively;
2. Assemble an integrated physical model by combining the kinetic model and the CFD model to simulate biomass growth and bioproduct synthesis under a few design scenarios (*i.e.* different operating conditions and PBR configurations);
3. Use the integrated model's simulation results (training data sets) to train a convolutional neural network (CNN);
4. Identify the desirable structure of the CNN to substitute the integrated physical model and predict system performance under new design conditions;
5. Determine the optimal PBR configurations and cultivation conditions for biomass growth and bioproduct synthesis through the hybrid stochastic optimization algorithm.
6. Validate the surrogate model predicted optimal solution via the integrated physical model. If this is unsatisfactory, return to Step 3.

In order to illustrate the above procedure and demonstrate the efficiency and accuracy of the modelling framework, a case study is provided with the aim to investigate simultaneously the optimal configurations of a 120 L flat-plate PBR (*i.e.* number and diameter of holes on the sparger) and the optimal batch operating conditions (*i.e.* incident light intensity and gas recycling rate) for the production of bisabolene, a novel sustainable biofuel synthesized from green microalga *Chlamydomonas reinhardtii* [8,23].

CFD modelling of a 120 L PBR

The 120 L pilot scale flat-plate PBR (Figure 1a) used in this study was purchased from Photon Systems Instruments (PSI). For the model, the system was divided into two separate compartments: the cuboidal growth chamber (1700 mm in length, 67 mm in width, 900 mm in depth) to hold liquid culture and microalgae biomass, and a sparger (20 mm in diameter, 1600 mm in length) located at the bottom of the cuboidal compartment through which recycling gas is pumped into the reactor providing essential mixing and carbon dioxide. As the configuration of the sparger has a critical influence on the liquid hydrodynamic behaviour of the flat-plate PBR, balancing the advantages of rapid liquid mixing with the potential risk of introducing high shear stress, in the current study both the number and diameter of holes on the sparger and the gas recycling rate are chosen as design variables to optimize the culture fluid dynamics.

Due to the small differences in liquid phase and algae biomass density, algal cell movement is approximated to be the bulk movement of the liquid culture. Hence, a gas-liquid multiphase computational fluid dynamic (CFD) simulation was carried out in the commercial software COMSOL Multiphysics 5.3®. The turbulent bubbly flow model was selected to

simulate the microalgal biomass cultivation compartment, and the single-phase flow model was used for the sparger [24]. Due to the significant scale difference of the cuboidal compartment (120 L) and the sparger holes (gas inlet, 1 mm to 5 mm), a customized tetrahedral meshing method was designed to generate 205,396 domain elements, 21,910 boundary elements, and 878 edge elements to guarantee the accuracy of the simulation.

The continuity equations and momentum balance equations for the gas-liquid multiphase model are shown in Eq. 1 and Eq. 2, respectively. Turbulent viscosity was solved using the standard *k*-ε model due to its widely successful application in the open literature [21,16]. For the single-phase sparger model, α is reduced to 1 and ϕ is reduced to 0 in the continuity equation, and the effective viscosity is replaced by gas viscosity in the momentum balance equation. The detailed CFD simulation procedure as well as the selection of boundary conditions can be found in our previous publication [24]. The current constructed 120 L PBR CFD model is shown in Fig. 1(b). The upper and lower bound of the three design variables (number and diameter of holes and gas inflow rate) are listed in Table 1. In this work, a 2^3 full factorial design was executed to generate 8 CFD models to simulate the fluid dynamics of the PBR under different sparger configuration and gas inflow rate.

$$\nabla \cdot (\rho\alpha\mathbf{u})_i + \frac{\partial(\rho\alpha)_i}{\partial t} = \phi_i \qquad 1(a)$$

$$\alpha_g + \alpha_i = 1 \qquad 1(b)$$

$$\phi_g + \phi_i = 0 \qquad 1(c)$$

$$\frac{\partial(\rho\mathbf{u})}{\partial t} + \nabla \cdot (\rho\mathbf{u}\mathbf{u}) = -\nabla p + \nabla \cdot \xi + \rho\mathbf{g} \qquad 2(a)$$

$$\boldsymbol{\xi} = \mu_e \left(\nabla \cdot \mathbf{u} + (\nabla \cdot \mathbf{u})^T - \frac{2}{3} I (\nabla \cdot \mathbf{u}) \right) \qquad 2(b)$$

where $\boldsymbol{\phi_i}$ is the interfacial mass transfer rate, ρ is the pseudo-continuous phase density, α is the volume fraction of phase, **u** is the velocity vector, i refers to different phases (g: gas phase, and l: liquid phase), **g** is the gravity acceleration vector, p is the pressure, $\boldsymbol{\xi}$ is the stress tensor and μ_e is the effective viscosity (including molecular viscosity and turbulent viscosity).

Kinetic modeling of algal biomass growth and bisabolene production

A kinetic model presented in Eq. 3 was developed in our recent work to simulate the effects of light intensity and cultivation temperature on algal biomass growth and bisabolene production under nutrient-sufficient conditions [23]. Equation 3a represents biomass growth rate, and its specific growth rate (μ) is a function of local light intensity and temperature formulated as Eq. 3b– Eq. 3d. To account for light attenuation [11], the modified Lambert-Beer law presented in Eq. 3e was adopted to calculate local light intensity. The kinetic model for parameter estimation was simplified to reduce its complexity (*i.e.* the model dimension) by replacing Eq. 3c with Eq. 3f to estimate an average value for $k(I)$ over the light transmission direction. Finally, Eq. 3g was constructed to simulate bisabolene production by modifying the Luedeking–Piret model [25]. The detailed explanation of this model and its parameter estimation method can be found in our recent work [23].

$$\frac{dX}{dt} = \mu \cdot X - \mu_d \cdot X^2 \qquad 3(a)$$

$$\mu = \mu_m \cdot k(I) \cdot k(T) \qquad 3(b)$$

$$k(I) = \frac{I(z)}{I(z) + k_s + \frac{I(z)^2}{k_i}} \qquad 3(c)$$

$$k(T) = \exp\left[-\left(\frac{E_a}{RT} - \frac{E_a}{RT_a}\right)\right] - \exp\left[-\left(\frac{E_b}{RT} - \frac{E_b}{RT_b}\right)\right] \qquad 3(d)$$

$$I(z) = I_0 \cdot \exp[-(\tau \cdot X + K_a) \cdot z] \qquad 3(e)$$

$$k(I) = \frac{1}{20} \cdot \sum_{n=1}^{9} \left(\frac{I_{i=0}}{I_{i=0} + k_s + \frac{I_{i=0}^2}{k_i}} + 2 \cdot \frac{I_{i=\frac{n \cdot L}{10}}}{I_{i=\frac{n \cdot L}{10}} + k_s + \frac{I_{i=\frac{n \cdot L}{10}}^2}{k_i}} + \frac{I_{i=L}}{I_{i=L} + k_s + \frac{I_{i=L}^2}{k_i}} \right) \quad 3(f)$$

$$\frac{dP}{dt} = \left(Y_1 \cdot \frac{dX}{dt} + Y_2 \cdot X \right) \cdot \left(\sigma - \left(\exp\left[-\left(\frac{E_a}{RT} - \frac{E_a}{RT_a} \right) \right] - \exp\left[-\left(\frac{E_b}{RT} - \frac{E_b}{RT_b} \right) \right] \right) \right) \quad 3(g)$$

where X is biomass concentration, μ is biomass specific growth rate, μ_d is biomass specific decay rate, μ_m is maximum specific growth rate, $I(z)$ and I_0 are local light intensity and incident light intensity, respectively, k_s and k_i are the photosaturation and photoinhibition term, respectively, E_a and E_b are the algae activation energy and deactivation energy, respectively, T is the culture temperature, T_a and T_b are reference temperatures, R is the gas constant (8.315 J mol^{-1} K^{-1}), τ is the algal cell absorption coefficient, K_a is the bubble scattering coefficient, z is the distance from light source, L is the width of the PBR, I_i is the local light intensity at a distance of $i = \frac{n \cdot L}{10}$ from the PBR exposure surface, P is bisabolene production, Y_1 and Y_2 are biomass growth-associated and growth-independent bisabolene yield coefficient, respectively, and σ is a temperature related dimensionless parameter.

Although both light intensity and temperature affect biomass growth and bioproduct synthesis, it is known that algal cell growth can be highly sensitive to small changes in culture temperature, particularly when operated beyond the optimal value [10]. This means that in practice, temperatures should be kept constant to maintain the activity of microalgae biomass for bioproduct synthesis, considering only light intensity as the design variable for biomass cultivation and process optimization in this study.

Integration of kinetic and CFD modelling

During the construction of the kinetic model, two simplifications were made to reduce the influence of hydrodynamics. Firstly, the culture was assumed to be perfectly mixed, and secondly, the local $k(I)$ (*i.e.* Eq. 3c was substituted with an averaged $k(I)$ value *i.e.* Eq. 3f). To integrate the kinetic model into the CFD models, both algal cells movement and local light distribution have to be included. Meanwhile, as temperature was fixed to 30 °C (suitable for biomass growth), $k(T)$ in Eq. 3d is reduced to a constant, σ'. Therefore, Eq. 4a and Eq. 4b were derived to replace Eq. 3a and Eq. 3g, respectively, and Eq. 3c was used directly to calculate local biomass growth (instead of Eq. 3f). While for lab scale PBRs, the effect of bubble scattering on light transmission was found to be negligible [6], it has been found to be an important factor on the local light distribution in large scale systems [26]. Thus, Eq. 3e was modified to Eq. 4c to ensure its applicability for the current 120 L PBR. The average bubble diameter was measured as 8 mm using video imaging.

$$\frac{dX}{dt} = \mu \cdot X - \mu_d \cdot X^2 - \nabla \cdot (\mathbf{u}X) \qquad 4(a)$$

$$\frac{dP}{dt} = (Y_1 \cdot (\mu \cdot X - \mu_d \cdot X^2) + Y_2 \cdot X) \cdot (\sigma - \sigma') - \nabla \cdot (\mathbf{u}P) \qquad 4(b)$$

$$I(z) = I_0 \cdot \exp\left[-\left(\tau \cdot X + \frac{3\alpha_g}{d_b}\right) \cdot z\right] \qquad 4(c)$$

Upon completion of the integrated models, they were applied to a batch operation, as this currently represents the most common operation mode for large scale biomass cultivation and biorenewables' production. The integrated models were performed until biomass concentration reached the stationary phase (maximum biomass concentration). They were then used to generate local values of biomass concentration, bisabolene production, and shear stress at different reactor locations. For the convenience of comparison, shear stress was converted to friction velocity using Eq. 5. In order to generate a wide range of samples for the surrogate model construction, incident light intensity was changed 5 times from 60 μE m^{-2} s^{-1}

to 300 μE m^{-2} s^{-1} (60 μE m^{-2} s^{-1}, 120 μE m^{-2} s^{-1}, 180 μE m^{-2} s^{-1}, 240 μE m^{-2} s^{-1}, 300 μE m^{-2} s^{-1}) in each configuration of the PBR to summarize cell growth and bisabolene production under different conditions. Table 1 lists the lower and upper bounds of incident light intensity.

$$u_f = \sqrt{\frac{\tau}{\rho_l}} \tag{5}$$

where τ is shear stress, ρ_l is liquid density, and u_f is friction velocity.

Consequently, a total of 40 scenarios were simulated under different process operating conditions (light intensity and recycling gas inflow rate) and photobioreactor configurations (number and diameter of holes on the sparger). For each scenario, approximately 9,000 data points (each point consists of biomass concentration, bisabolene production, and friction velocity at a specific location) were generated, resulting in 360,000 data points for the surrogate model construction.

Construction of surrogate model

Convolutional Neural Networks (CNNs) are a variant of Artificial Neural Networks (ANNs) of the multi-layer Perceptron type. The following is a highly simplified explanation, and interested readers should be pointed out to highly informative reviews [27,28]. CNNs have different types of layers compared to ANNs. In a CNN, an input first meets one or more convolutional layers, where a number of “Filters” or “Kernels” steps over the input tensor. These detect features, as an example they might detect vertical or horizontal line in an image, with filters in deeper layers detecting progressively more complex features. For instance, a filter in a deep network might have learned to look for a window or mailbox, if the network was trained to detect houses. The output then passes first through a function such as RELU or sigmoid, then a pooling layer, which reduces the size of the output while preserving

information. This is completed by extracting the strongest activations, and doing so also means that the network is less susceptible to small changes and overfitting. Convolutional and pooling operations are sometimes considered a single layer. At the end of the CNN there is one or more fully connected layers of linked neurons producing readable predictions, similar to traditional ANNs.

As pointed out in [28], CNNs possess highly useful properties. Parameter sharing is the first, meaning that a filter used for a convolution is used all over the input, unlike weights in a traditional ANN which are each used only on a fraction of the input variables. Another connected property is that of sparse weights, meaning that filters only require very few elements of their tensors to be non-zero. Both of these properties reduce the memory requirements for CNNs and the amount of fine-tuning necessary to find a working solution. The property of equivariance means that the result of a convolution on a slightly altered output is the same as if the alteration was executed on the original output, guaranteeing consistency in the output. This relatively light-weight structure and capacity to discriminate make them highly useful in pattern-spotting tasks. Furthermore, the ability to use “transfer learning” means that networks can be pre-trained to learn on similar tasks, expanding the amount of available data [29].

Therefore, in this study, a three-layer deep CNN model was constructed with 2 hidden layers consisting of convolutional blocks and a fully connected linear activation function output layer. A CNN was chosen deemed particularly suitable for the problem at hand. The first reason is that its pattern detection capabilities tolerate noise and uncertainty, allowing to understand general trends and to simulate robustly the system gradient landscape for the next step in the framework. The second is that this system is relatively slow changing, with many

readings being close to each other except for instrumentation noise. This might cause overfitting in another type of machine learning tool, but CNNs are resistant in this sense. The number of neurons (inputs) was 7 in the input layer (corresponding to incident light intensity, number of holes, diameter of holes, recycling gas flow rate, and cardinal coordinates x, y and z for each sample), 21 in each hidden layer and 3 in the output layer (corresponding to biomass concentration, friction velocity and bisabolene production). The number of neurons in the hidden layer was chosen as a simple multiple of existing inputs and outputs. Through early tuning, it was found that increasing the number of neurons did not significantly increase the accuracy of the model, while increasing the computational load.

For each layer in the network, the parameters inside each neuron were initialized using a truncated normal distribution multiplied by the square root of the inverse of the number of inputs that feature in the neuron. The network was trained for 100 iterations using a mini-batch size of 1024, a size that was found to be a good compromise between computational performance and training quality. A learning rate was given with an exponential decay of 0.98 applied every 7 iterations, starting from a value of 0.01. The generated data set was separated into two splits, one comprising 70% used for the network training and the other containing 30% used for network cross validation (test), a common split ratio employed in this type of work. The test points were chosen randomly by a Sobol sequence from the original data set. The implementation of the CNN based surrogate model construction was executed in a Python 3.6 programming environment using *TensorFlow* (Dean and Monga, 2015) and the Adam optimization algorithm (Kingma and Ba, 2015) on an Intel Core i7 2.40 GHz 16GB RAM Alienware laptop computer.

Stochastic optimization of the surrogate model

Once the surrogate model was constructed, an optimization was conducted to find the optimal reactor configurations and operating conditions with respect to a specific objective. Given that the CNN structure is highly nonlinear, multimodal and coupled with a mixed-integer nonlinear programming (MINLP) problem, gradient based optimization approaches might not be effective to find the optimal solutions. Hence, a hybrid stochastic search optimization algorithm is implemented in this work.

This hybrid algorithm uses a number *N* of particles with coordinates (the four design variables) as its agent and searches the solution space for the optimal result of a specific index through a combined form as: a) Random Search (*RS*), b) Particle Swarm Optimization (*PSO*) and c) Simulated Annealing (*SA*). Furthermore, when the algorithm detects that there are no good solutions outside a specific domain, the search space is reduced to focus on the regions which are more likely to contain a high-quality solution. To measure the quality of the different agents (design variables), three different performance indices (objective functions) were defined during the optimization process, which are biomass concentration, bisabolene production, and a combination of both. The definition of these performance indices are explained as follows.

COMSOL Multiphysics 5.3® outputs a grid of solution points as simulation results. This grid is the state values (biomass concentration, bisabolene production and friction velocity) at different locations of the 3-dimensional space inside the reactor. Henceforth, the CNN also outputs predictions on the state values at these locations. Given that the performance index should measure biomass concentration and bisabolene production inside of the reactor, the objective function with respect to the state *S* (biomass concentration or bisabolene production) is defined as:

$$obj_{MO} = \frac{1}{NP}\sum_{i=1}^{NP} S_i \quad (6)$$

where *NP* is the number of points computed by COMSOL, and S_i is the value for the i^{th} point computed by the CNN for state *S*. In this way, the average bioproduct concentration/production is approximated by the sum of punctual concentrations/productions inside the reactor, divided by the number of points. For the current implementation *NP* was set to 280 equally distributed points in the 3D reactor space, as the addition of further points had a minimal impact on the states' average.

In addition to Eq. 6, the maximum friction velocity at every computed point was restricted to be lower than 0.5 cm s^{-1}; otherwise the solution was discarded. This corresponds to the previous study in which friction velocity higher than 0.5 cm s^{-1} was found to damage the cell activities of *C. reinhardtii* cells [18]. The optimization was conducted using a 5-step optimization algorithm proposed in our earlier work [30] and consisting of the following steps:

Algorithm 1: hybrid stochastic optimization algorithm

I. Initialization: In this phase, *N* particles are randomly placed in the solution space of the optimization problem.

II. Evaluation and Classification: In this phase, the performance criteria of each particle are evaluated and classified into three groups. The group that they are assigned to will determine the search strategy they will follow in subsequent steps of the algorithm.

   i. A number N_{SA} of particles with the highest evaluated performance index will follow a *SA* search strategy for the following *n* iterations.

   ii. A number N_{RS} of particles with the lowest evaluated function will follow a *RS* search strategy for the following *n* iterations.

iii. The remaining N_{PSO} particles will follow a *PSO* search strategy for the following *n* iterations.

III. Space Exploration: The above classification works in such a way that the position of each particle is better exploited depending on the information that it can supply to the swarm.

i. The N_{SA} particles should be a small number of particles which are assumed to be near high-quality solutions. Thus, their neighbourhood will be intensively explored.

ii. A similar reason is applied to the N_{PSO} particles. In the solution space, these particles are positioned such that they might not be close to high-quality solutions. However, by using the knowledge of other particles in the swarm they can search for better solution areas which have not yet been found by other particles.

iii. Finally, the N_{RS} particles are those assumed to be relatively far away from any high-quality solutions. Hence, exploring the neighbourhood around them would not bring benefit to the swarm. Then, particles classified as N_{RS} are reinitialised randomly.

IV. Repeat: After *n* iterations the algorithm either returns to the Evaluation and Classification phase or terminates.

V. Space reduction: During the algorithm, the best position of each particle in the current cycle is recorded. After N_{cycle} iterations, the search space is reduced to the smallest hypercube that contains the best position of all particles in the current cycle, this terminates the cycle. Subsequently, the algorithm returns to the Initialization phase.

In this way, given that RS, PSO, and SA are in ascending order of exploration and in descending order of exploitation a balance in the exploration-exploitation paradigm can be

achieved [30]. Furthermore, the space reduction allows the algorithm to focus efforts in the areas that are most likely to have a high quality (or possibly the global) optimum.

Using the above optimization algorithm, optimal solutions for biomass cultivation and bisabolene production were sought. After observing that the optimal solutions for the two indices were distinct from each other, a multi-objective optimization framework was implemented. The objective indices for biomass concentration and biofuel production were scaled with respect to their optimum values, and a new performance index was created by merging both previous objective functions. Subsequently, using the Weighted Sum Method a Pareto frontier was obtained to determine the compromise point [31]. The objective function used in this case was Eq. 7:

$$obj_{MO} = \frac{1}{NP}\left(\theta \sum_{i=1}^{NP} biomass_i + (1-\theta)\sum_{i=1}^{NP} bisabolene_i\right) \quad (7)$$

where θ is a parameter running from 1 to 0 computing different values for the trade-off between biomass concentration and bisabolene production inside the reactor.

The compromise point was defined to be the closest point in the Pareto front to the utopia point, given the Euclidean distance. The implementation was performed in a Python 3.6 programming environment using the *numpy* library for fast vectorized implementations on an Intel Core i7 2.40 GHz 16GB RAM Alienware laptop computer.

Results and Discussion

Results of the surrogate model

In the current study, 70% of data points generated from the 40 scenarios were used to construct the CNN based surrogate model, meaning that 252,000 data points were selected to

train the convolutional CNN. The training time course was around 2 hours, negligible compared to the time spent on converging a CFD model (average 144 hours). To test the accuracy of the surrogate model, the remaining 108,000 data points were used for CNN cross-validation, and the results are presented in Table 2. From the table, it is seen that the surrogate model can well represent the complex fluid dynamics and process kinetics simulated by the integrated kinetic-CFD model. Therefore, it was used to replace the rigorous physical model given its superior efficiency in numerical calculations and mathematical optimization, and was applied to predict untested behaviours of the system throughout a large solution space of design variables to seek the optimal solution for further PBR design and batch operation. To verify its predictive capability, the identified optimal PBR configuration and operating conditions were passed to the physical model to generate the "authentic" result for final comparison.

Single-objective optimization for biomass growth and bisabolene production

The optimized sparger configurations and operating conditions obtained by the surrogate model for biomass cultivation and bisabolene production are listed in Table 3. The baseline results listed in Table 3 refer to the highest values (1.027 g L^{-1} biomass concentration and 50.77 μg L^{-1} bisabolene production) found in the 40 scenarios used for training data generation. From the table, it can be seen that through optimization, the current study can result in an increase up to 15% for different objective indices. Comparison between the CNN prediction result and the integrated model verification result is presented in Table 4. From Table 4, it is seen that the surrogate model possesses great predictive capabilities for both biomass concentration and biofuel production, with a mean error below 1% in both cases. Although the friction velocity is underestimated by the CNN, the real value (calculated by

the integrated model) for both cases remains significantly below the limiting velocity of 0.5 cm s^{-1}, and is therefore to have no adverse effect on the viability of *C. reinhardtii* cells [18].

The optimization results (Table 3) show that despite the similarity in gas inflow rate and number of holes in the sparger, the incident light intensity for optimized bisabolene production is less than half that for optimized biomass growth, and the hole diameter drops from 5 mm (upper bound) to 1 mm (lower bound). These differences can be attributed to the different kinetic mechanisms of biomass growth and bisabolene production and the already mentioned decreasing light intensity along the light transmission direction within the PBR. Based on the available photons that can be absorbed by cells, the culture can be divided into two zones, a light zone, where cells receive sufficient energy of photons to conduct photosynthesis, and a dark zone, where photosynthesis is no longer supported and cells consume their intracellular storage compounds to maintain their metabolic activities resulting in cell decay [26].

To maximize biomass concentration, liquid mixing needs to be intense enough for cells to frequently enter the light zone to enable growth and maintenance. Increasing the diameter of holes on a sparger was previously found to rearrange the distribution of local bubble volume fractions, effectively enhancing liquid movement [24]. Therefore, the optimal hole diameter for maximum biomass growth was predicted at the upper bound value of 5 mm. Meanwhile, increasing the incident light intensity also increases the local light intensities within the 120 L PBR, maximizing the volume of the light zone. Thus, the optimal incident light intensity is found to be 300 µE m^{-2} s^{-1} for the biomass production scenario.

In contrast, bisabolene synthesis is not primarily determined by biomass growth, but occurs even after the biomass reaches the stationary phase [23]. Moreover, although bisabolene is produced during the cell growth phase as well, its synthesis rate was found to decrease at conditions favouring cell growth. At optimal conditions for algae biomass growth, cells will direct the majority of carbon towards their own reproduction, instead of synthesising other secondary metabolites, a phenomenon frequently found by other research *e.g.* biolipid production [3,32]. As a result, to achieve maximum bisabolene production, biomass concentration should be controlled at a level which provides sufficient cells for bisabolene synthesis, while maintaining low growth activity. Consequently, both the hole diameter and incident light intensity are decreased for this scenario. From Table 4, it is concluded that in this case bisabolene productivity (μg g^{-1}) is improved by 22.1 % compared to the scenario of maximizing biomass growth.

Multi-objective optimization for biomass growth and bisabolene production

To balance the trade-off between microalgal cell growth and biofuel production, multi-objective optimization was implemented, and a Pareto frontier was generated as shown in Fig. 2. The untraditional frontier further highlights the high nonlinearity and complexity of the current optimization problem, thus indicating the competence of the proposed modelling framework. Based on the Pareto frontier, the compromise point was identified and presented in Table 3, while the corresponding biomass concentration and bisabolene production are listed in Table 4. The optimal bisabolene productivity was estimated to be 48.55 μg g^{-1}. Similarly to the previous two scenarios, the surrogate model can well predict biomass concentration and bisabolene production, but still underestimates the friction velocity. Nonetheless, based on the integrated model, the friction velocity remains below 0.5 cm s^{-1}, and therefore does not restrain biomass growth.

Advantages of the current modelling framework

Finally, to examine further the accuracy and predictive capability of the current surrogate model, an analysis was carried out to compare local biomass concentration and bisabolene production between the surrogate model prediction result and integrated model verification result for all three cases. Based on the analysis, it is concluded that for all cases, the prediction error of local biomass concentration does not exceed 2.5% throughout 6,000 samples at different locations in the pilot scale PBR, while the error for local biofuel production is less than 1.0%. An example is also presented in Figure 3. This conclusion is strong evidence for the accuracy and predictive power of the proposed modeling framework.

A major advantage of the current modelling framework is that it only requires the simulation result of a few design scenarios (*i.e.* two values for each design variable, in total 8 scenarios) from the physical model which in average requires 144 hours (6 days) to converge for each configuration of the pilot-scale PBR. This leads to a negligible overall calculating time compared to the traditional approach. For example, the total time spent in this work was around 7 weeks, including approx. 1,152 hours to run a few CFD models (each representing one configuration of PBR) and less than 1 day to execute the surrogate modelling framework. As the estimation of the process' biochemical behaviors under each configuration of PBR took insignificant time to converge (less than 1 hour), it has been omitted in this estimate.; In contrast, current methods (*e.g.* response surface methodology (RSM)) require at least three levels of each design variable (*i.e.* 27 design scenarios) to get a good regression for optimization purpose [33]. As a result, in total 23 weeks (3,864 hours) will be spent in the entire modelling procedure. Hence, the proposed surrogate modeling framework can greatly reduce computing time from months (16 weeks) to days (1 day), meanwhile guaranteeing a high

accuracy of the optimization result. Therefore, the proposed surrogate modelling framework may represent the only feasible solution for the rigorous design and optimization of large scale bioreactors if there is a larger number of design variables that have to be taken into account.

More importantly, RSM and other regression methods tend to use a simple formulation *e.g.* quadratic equation to approximate the relation between the target index and the design variables [33]. However, the current bioprocess is governed by sophisticated fluid dynamics and biochemical kinetics, indicating that its behaviour is complex and cannot be oversimplified by using a quadratic equation or other similar formulations. . Thus, neither the accuracy nor the reliability can be assured when using RSM to seek optimal photobioreactor configuration and process operating conditions. Based on this comparison, it is concluded that the current modelling framework represents a more accurate and efficient strategy, particularly for the modelling of complex multi-scale biosystems.

Conclusion

In the current study, a surrogate modelling framework was proposed to simultaneously identify the optimal configurations and operating conditions for pilot scale photobioreactor design and sustainable microalgal biofuel production. To guarantee the accuracy of the framework, an integrated kinetic-CFD model was initially constructed to investigate the behaviours of the underlying biosystem and to generate a sufficient number of data sets. Then, a convolutional artificial neuron network was developed through a state-of-the-art structure selection approach so that it can well represent the high nonlinearity and complexity of the original models. By implementing the robust hybrid stochastic optimization algorithm, optimal solutions with respect to different indices were successfully identified.

From these results it was found that due to the intricate hydrodynamic and biochemical mechanisms, the optimal incident light intensity and number of holes on the sparger decrease remarkably when the objective switches from biomass cultivation to biofuel production. This results in a 10% decrease on final biomass concentration but 22% increase on bisabolene productivity, which clearly suggests the importance of coupling effects of fluid dynamics and biological kinetics for bioprocess optimization, and the necessity of developing an efficient multi-scale modelling approach for bioprocess integration. Moreover, multi-objective optimization was executed to balance both cell growth and biofuel production. In all cases, friction velocity was well controlled to minimize the culture dead zone. All the three optimization schemes also resulted in noticeable increases in their objective indices compared to the initial design scenarios. Finally, from this detailed analysis, the current framework was demonstrated to combine great predictive capability with high computational efficiency, indicating its applicability for general biosystems modelling and optimization.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 640720. This project has also received funding from the EPSRC project (EP/P016650/1, P65332).

Literature Cited

1. Wagner JL, Le CD, Ting VP, Chuck CJ. Design and operation of an inexpensive, laboratory-scale, continuous hydrothermal liquefaction reactor for the conversion of microalgae produced during wastewater treatment. *Fuel Process Technol.* 2017;165:102-111. doi:10.1016/j.fuproc.2017.05.006
2. Slade R, Bauen A. Micro-algae cultivation for biofuels: Cost, energy balance,

environmental impacts and future prospects. *Biomass and Bioenergy*. 2013;53:29-38. doi:10.1016/j.biombioe.2012.12.019

3. Cakmak T, Angun P, Demiray YE, Ozkan AD, Elibol Z, Tekinay T. Differential effects of nitrogen and sulfur deprivation on growth and biodiesel feedstock production of Chlamydomonas reinhardtii. *Biotechnol Bioeng*. 2012;109(8):1947-1957. doi:10.1002/bit.24474
4. Zhang D, Vassiliadis VS. Chlamydomonas reinhardtii Metabolic Pathway Analysis for Biohydrogen Production under Non-Steady-State Operation. *Ind Eng Chem Res*. 2015;54(43):10593-10605. doi:10.1021/acs.iecr.5b02034
5. Giannelli L, Yamada H, Katsuda T, Yamaji H. Effects of temperature on the astaxanthin productivity and light harvesting characteristics of the green alga Haematococcus pluvialis. *J Biosci Bioeng*. 2014;xx(xx). doi:10.1016/j.jbiosc.2014.09.002
6. del Rio-Chanona EA, Ahmed N rashid, Zhang D, Lu Y, Jing K. Kinetic modeling and process analysis for Desmodesmus sp. lutein photo-production. *AIChE J*. 2017;63(7):2546-2554. doi:10.1002/aic.15667
7. Davies FK, Work VH, Beliaev AS, Posewitz MC. Engineering Limonene and Bisabolene Production in Wild Type and a Glycogen-Deficient Mutant of Synechococcus sp. PCC 7002. *Front Bioeng Biotechnol*. 2014;2. doi:10.3389/fbioe.2014.00021
8. Wichmann J, Baier T, Wentnagel E, Lauersen KJ, Kruse O. Tailored carbon partitioning for phototrophic production of (E)-α-bisabolene from the green microalga Chlamydomonas reinhardtii. *Metab Eng*. 2018;45:211-222. doi:10.1016/j.ymben.2017.12.010
9. Xie Y, Jin Y, Zeng X, Chen J, Lu Y, Jing K. Fed-batch strategy for enhancing cell

growth and C-phycocyanin production of Arthrospira (Spirulina) platensis under phototrophic cultivation. *Bioresour Technol*. 2015;180:281-287. doi:10.1016/j.biortech.2014.12.073

10. Zhang D, Dechatiwongse P, del Rio-Chanona EA, Maitland GC, Hellgardt K, Vassiliadis VS. Modelling of light and temperature influences on cyanobacterial growth and biohydrogen production. *Algal Res*. 2015;9:263-274. doi:10.1016/j.algal.2015.03.015
11. del Rio-Chanona EA, Liu J, Wagner JL, et al. Dynamic modeling of green algae cultivation in a photobioreactor for sustainable biodiesel production. *Biotechnol Bioeng*. 2018;115(2):359-370. doi:10.1002/bit.26483
12. Berberoglu H, Yin J, Pilon L. Light transfer in bubble sparged photobioreactors for H2 production and CO2 mitigation. *Int J Hydrogen Energy*. 2007;32(13):2273-2285. doi:10.1016/j.ijhydene.2007.02.018
13. Posten C. Design principles of photo-bioreactors for cultivation of microalgae. *Eng Life Sci*. 2009;9(3):165-177. doi:10.1002/elsc.200900003
14. del Rio-Chanona EA, Manirafasha E, Zhang D, Yue Q, Jing K. Dynamic modeling and optimization of cyanobacterial C-phycocyanin production process by artificial neural network. *Algal Res*. 2016;13:7-15. doi:10.1016/j.algal.2015.11.004
15. Adesanya VO, Davey MP, Scott SA, Smith AG. Kinetic modelling of growth and storage molecule production in microalgae under mixotrophic and autotrophic conditions. *Bioresour Technol*. 2014;157:293-304. doi:10.1016/j.biortech.2014.01.032
16. Bitog JP, Lee I-B, Lee C-G, et al. Application of computational fluid dynamics for modeling and designing photobioreactors for microalgae production: A review. *Comput Electron Agric*. 2011;76(2):131-147. doi:10.1016/j.compag.2011.01.015
17. Ugwu CU, Aoyagi H, Uchiyama H. Photobioreactors for mass cultivation of algae.

Bioresour Technol. 2008;99(10):4021-4028. doi:10.1016/j.biortech.2007.01.046

18. Leupold M, Hindersin S, Gust G, Kerner M, Hanelt D. Influence of mixing and shear stress on Chlorella vulgaris, Scenedesmus obliquus, and Chlamydomonas reinhardtii. *J Appl Phycol*. 2013;25(2):485-495. doi:10.1007/s10811-012-9882-5
19. Sforza E, Simionato D, Giacometti GM, Bertucco A, Morosinotto T. Adjusted Light and Dark Cycles Can Optimize Photosynthetic Efficiency in Algae Growing in Photobioreactors. Webber A, ed. *PLoS One*. 2012;7(6):e38975. doi:10.1371/journal.pone.0038975
20. Michels MHA, van der Goot AJ, Vermuë MH, Wijffels RH. Cultivation of shear stress sensitive and tolerant microalgal species in a tubular photobioreactor equipped with a centrifugal pump. *J Appl Phycol*. 2016;28(1):53-62. doi:10.1007/s10811-015-0559-8
21. Park S, Li Y. Integration of biological kinetics and computational fluid dynamics to model the growth of Nannochloropsis salina in an open channel raceway. *Biotechnol Bioeng*. 2015;112(5):923-933. doi:10.1002/bit.25509
22. Nikolaou A, Booth P, Gordon F, Yang J, Matar O, Chachuat B. Multi-Physics Modeling of Light-Limited Microalgae Growth in Raceway Ponds. *IFAC-PapersOnLine*. 2016;49(26):324-329. doi:10.1016/j.ifacol.2016.12.147
23. Harun I, Del Rio-Chanona EA, Wagner JL, Lauersen KJ, Zhang D, Hellgardt K. Photocatalytic Production of Bisabolene from Green Microalgae Mutant: Process Analysis and Kinetic Modeling. *Ind Eng Chem Res*. 2018;57(31):10336-10344. doi:10.1021/acs.iecr.8b02509
24. Zhang D, Dechatiwongse P, Hellgardt K. Modelling light transmission, cyanobacterial growth kinetics and fluid dynamics in a laboratory scale multiphase photo-bioreactor for biological hydrogen production. *Algal Res*. 2015;8:99-107. doi:10.1016/j.algal.2015.01.006

25. Obeid J, Flaus J-M, Adrot O, Magnin J-P, Willison JC. State estimation of a batch hydrogen production process using the photosynthetic bacteria Rhodobacter capsulatus. *Int J Hydrogen Energy*. 2010;35(19):10719-10724. doi:10.1016/j.ijhydene.2010.02.051

26. Huang Q, Jiang F, Wang L, Yang C. Design of Photobioreactors for Mass Cultivation of Photosynthetic Organisms. *Engineering*. 2017;3(3):318-329. doi:10.1016/J.ENG.2017.03.020

27. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. *Neurocomputing*. 2017;234:11-26. doi:10.1016/j.neucom.2016.12.038

28. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks*. 2015;61:85-117. doi:10.1016/j.neunet.2014.09.003

29. Oquab M, Bottou L, Laptev I, Sivic J. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2014:1717-1724. doi:10.1109/CVPR.2014.222

30. Estrada-Wiese D, del Río-Chanona EA, del Río JA. Stochastic optimization of broadband reflecting photonic structures. *Sci Rep*. 2018;8(1):1193. doi:10.1038/s41598-018-19613-6

31. Miettinen K. *Nonlinear Multiobjective Optimization*. Vol 12. 1st ed. Boston, MA: Springer US; 1998. doi:10.1007/978-1-4615-5563-6

32. Mairet F, Bernard O, Masci P, Lacour T, Sciandra A. Modelling neutral lipid production by the microalga Isochrysis aff. galbana under nitrogen limitation. *Bioresour Technol*. 2011;102(1):142-149. doi:10.1016/j.biortech.2010.06.138

33. Bezerra MA, Santelli RE, Oliveira EP, Villar LS, Escaleira LA. Response surface

methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*. 2008;76(5):965-977. doi:10.1016/j.talanta.2008.05.019

Table 1: Boundary values of the design variables. The range of number of holes was chosen based on the current PBR configuration; the total gas inflow rate (gas pumped from both sides of the sparger) and incident light intensity were determined by the working range of the 120 L photobioreactor.

Design variables	Lower bound	Upper bound
Diameter of holes	1 mm	5 mm
Number of holes	62	122
Gas inflow rate	5 L min^{-1}	20 L min^{-1}
Incident light intensity	60 µE m^{-2} s^{-1}	300 µE m^{-2} s^{-1}

Table 2: Training result of the surrogate model

Model output	Mean error, %	Standard error, %	Maximum error, %
Biomass concentration	0.21%	1.07%	14.7%
Bisabolene production	0.19%	3.75%	16.1%
Friction velocity	1.22%	2.45%	17.5%

Table 3: Optimal configurations and operating conditions under different objective functions. The baseline results refer to the highest values (1.027 g L^{-1} biomass concentration and 50.77 µg L^{-1} bisabolene production) found in the 40 scenarios in this work.

	Optimize biomass	Optimize bisabolene	Multi-objective Opt.
Incident light intensity	300 µE m^{-2} s^{-1}	142 µE m^{-2} s^{-1}	214 µE m^{-2} s^{-1}
Number of holes	84	118	98
Diameter of holes	5 mm	1 mm	5 mm
Gas inflow rate	18 L min^{-1}	14 L min^{-1}	20 L min^{-1}
Increase compared to the baseline results	14.3% (biomass)	13.5% (bisabolene)	12.1% (biomass) 13.8% (bisabolene)

Table 4: Comparison between the surrogate model prediction result (S-Model prediction) and integrated kinetic-CFD model verification result (K-CFD verification)

Single-objective function: Maximize final biomass concentration			
Output variables	S-Model prediction	K-CFD verification	Deviation
Biomass concentration	1.174 g L^{-1}	1.174 g L^{-1}	0.0%
Bisabolene production	51.82 µg L^{-1}	51.83 µg L^{-1}	0.006%
Bisabolene productivity	44.13 µg g^{-1}	44.15 µg g^{-1}	0.05%
Friction velocity	0.072 cm s^{-1}	0.079 cm s^{-1}	9.62%
Single-objective function: Maximize final bisabolene production			
Output variables	S-Model prediction	K-CFD verification	Deviation
Biomass concentration	1.069 g L^{-1}	1.069 g L^{-1}	0.0%
Bisabolene production	57.59 µg L^{-1}	57.61 µg L^{-1}	0.027%
Bisabolene productivity	53.87 µg g^{-1}	53.89 µg g^{-1}	0.04%
Friction velocity	0.128 cm s^{-1}	0.169 cm s^{-1}	32.5%
Multi-objective function: Maximize biomass growth and bisabolene production			
Output variables	S-Model prediction	K-CFD verification	Deviation
Biomass concentration	1.151 g L^{-1}	1.148 g L^{-1}	0.258%
Bisabolene production	57.76 µg L^{-1}	55.73 µg L^{-1}	0.064%
Bisabolene productivity	50.18 µg g^{-1}	48.55 µg g^{-1}	3.36%
Friction velocity	0.134 cm s^{-1}	0.201 cm s^{-1}	50.0%

List of Figure Captions:

Figure 1: The 120 L flat-plate photobioreactor (a) and its CFD model (b).

Figure 2: Pareto frontier of the multi-objective optimization problem.

Figure 3: Prediction error of the surrogate model in case 1 (optimizing biomass concentration). (a): Prediction error on biomass concentration; (b): Prediction error on bisabolene production.