Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Fed-batch bioprocess prediction and dynamic optimization from hybrid modelling and transfer learning

Oliver Pennington [a], Youping Xie [b], Keju Jing [c], Dongda Zhang [a]*

[a] University of Manchester, Department of Chemical Engineering, Manchester, United Kingdom
[b] Fuzhou University, Marine Biological Manufacturing Center or Fuzhou Institute of Oceanography, Fuzhou, China
[c] Xiamen University, Department of Chemical and Biochemical Engineering, College of Chemistry and Chemical Engineering, Xiamen, China
* Corresponding Author: dongda.zhang@manchester.ac.uk.

ABSTRACT

Hybrid modelling utilizes advantageous aspects of both mechanistic (white box) and data-driven (black box) modelling. Combining the physical interpretability of kinetic modelling with the power of a data-driven Artificial Neural Network (ANN) yields a hybrid (grey box) model with superior accuracy when compared to a traditional mechanistic model, while requiring less data than a purely data-driven model. This study demonstrates the construction a hybrid model with transfer learning for the predictive modelling and optimization of a high-cell-density microalgal fermentation process for lutein production. Dynamic optimization was conducted to identify a feeding strategy that maximized final lutein production. The results were then experimentally validated. Overall, this work presents a novel digital twin application that can be easily adapted to general bioprocesses for model predictive control and process optimization.

Keywords: Machine Learning, Dynamic Modelling, Biosystems, Hybrid Modelling, Dynamic Optimization

1 INTRODUCTION

1.1 Motivation

Bioprocesses connect an abundance of global research interests, including the production of renewable plastics, fuels, and other valuable bioproducts. Economically speaking, the UK bioeconomy alone was worth around £220 billion in 2018, and is set to double by 2030 [1], making the field lucrative for research. Supporting bioprocess development requires overcoming several challenges, including batch-to-batch variation, deficient metabolic and secretory phenotypes for protein production, by-product accumulation, and reduced yields in the upscaling of bioreactors. Tackling these challenges involves identifying suitable microbial strains and their optimal operating conditions, as well as the development of optimal operation and control strategies. Conducting such process optimization exhaustively can be extremely time-consuming and resource-depleting; employing modelling can reduce the time and resources spent by using predictions to rule out less useful experiments.

With the ever-progressing fourth industrial revolution underway, digitalization is seeing increased application; utilizing digital twins has significant potential in the application of process optimization and control, as well as design of experiments. The concept of machine learning is being continuously applied in novel ways to unveil more economical, ethical, safer, efficient, and sustainable approaches to chemical processing. With the current growth of interest within the fields of both bioprocess engineering and artificial intelligence, there has not been a better time to harness and combine the advantages of each. However, accounting for uncertainty is an equally important step in the development of a model as it provides information surrounding the achievability of a found control strategy or optimum.

An example of where the combining of different modelling methodologies may be required is high-cell-density cultures (HCDCs). HCDCs can be capable of producing larger quantities of desired products, an example of which is lutein, a carotenoid pigment commonly used in the food industry. However, HCDCs can demonstrate self-inhibitory effects, where growth and/or substrate uptake is inhibited by the high cell density or high accumulation of products [2]. Gaining a better understanding

of these inhibitory effects is pertinent for optimizing processes utilizing HCDCs.

1.2 Aims

Ultimately, this model aspires to require minimal data and computational expense, while maximizing simulation accuracy for process optimization and control applications. The model should also have good generalisability, allowing for easy incorporation of new data to retrain the model. Fulfilling these goals will make the modelling methodology described in this work applicable to other bioprocesses involving HCDCs, such as Akkermansia muciniphila which has received recent attention for its probiotic capabilities in the intestinal tract [3]. With HCDCs being well suited to fed-batch and perfusion operation, they play a key role in moving towards the continuous operation of bioprocesses; such progression can be accelerated by the modelling and optimization strategy discussed in the recent study [4].

In this work, the term *hybrid modelling* relates to the combination of a physically derivable (white-box) model, with a data-driven (black-box) model to yield an overall hybrid (grey-box) model.

1.3 Case Study

The case study at hand looks at the production of lutein from Chlorella sorokiniana under fed-batch conditions using urea as the source of nitrogen and glucose as the carbon source. The dataset available also includes a preliminary batch using sodium nitrate (an alternative to urea) as the nitrogen source. Experiments were run for 168 hours with measurements of biomass density and lutein content taken art least every 12 hours. The fed-batch saw continuous feeding of high concentrations of glucose and urea (750 g L^{-1} and 46.5 g L^{-1}, respectively).

2 METHODOLOGY

2.1 Experimental setup

Initially, a single batch process in a 250 mL flask (with 100 mL of working volume) was conducted using $NaNO_3$ as the nitrogen source, to which a preliminary kinetic model was fitted. Fed-batch operation was then conducted in a 5 L bioreactor with urea (which behaves similarly to $NaNO_3$) as the nitrogen source and continuous feeding over 168 h.

2.2 Macro-scale kinetic modelling

For the initial batch experiment, Monod-inspired kinetics were utilized. The concentrations of the substrates glucose (G) and urea (N), as well as cell density (X) and lutein content (L_C), are described by the system of 4 Ordinary Differential Equations (ODEs) as in

$$\frac{dX}{dt} = \mu \cdot X - d_X \cdot X \tag{1}$$

$$\frac{dG}{dt} = -v_{\max_G} \cdot \frac{G}{K_G+G} \cdot X \tag{2}$$

$$\frac{dN}{dt} = -v_{\max_N} \cdot \frac{N}{K_N+N} \cdot X \tag{3}$$

$$\frac{dL_C}{dt} = Y_{LX} - d_L \cdot L_C - \mu \cdot L_C \tag{4}$$

$$\mu = \mu_{\max} \cdot \frac{G}{K_G+G} \tag{5}$$

where $\mu_{\max}$ refers to the maximum specific growth rate of biomass, with $v_{\max_i}$ being the maximum specific uptake rate for a given substrate i, K_i being the affinity constant for a given substrate i, Y_{LX} being the specific production rate of lutein from biomass, and d_i representing a decay term for a given product i. It should be noted that L_C represents an intracellular concentration measured in mg (g biomass)$^{-1}$. To consider the overall lutein concentration in the reactor (L), product rule can be used with equations (1) and (4) to give the result as in

$$\frac{dL}{dt} = Y_{LX} \cdot X - d_L \cdot L \tag{6}$$

It is essential for parameters in the Monod model to remain positive in value to retain physical feasibility, as in

$$\boldsymbol{\beta} \geq \mathbf{0} \tag{7}$$

where $\boldsymbol{\beta}$ is the vector of Monod model parameters. Parameters were identified by minimizing the mean squared error objective function, as in

$$\min \frac{1}{n} \cdot \sum_{t,\text{meas}} \sum_i \left(\frac{C_{i_t} - C_{i_{\text{meas}_t}}}{\sigma_i} \right)^2 \tag{8}$$

where n is the number of datapoints. The objective function minimizes the difference between the simulated extracellular concentration profiles and the measured averages at each timepoint; each term i is weighted by its experimental measurement standard deviation; σ_i.

2.3 Hybrid modelling and transfer learning

Despite good fitting of the initial batch process, the parameters identified could not capture the entire fed-batch process, highlighting the difference between the $NaNO_3$-fed and urea-fed systems. The most noticeable process-model mismatches can be seen later in the process during high cell densities, particularly once the culture reaches a cell density over 5 times that seen in the preliminary batch experiment (seen around 30 h into the fed-operation). This indicates that substrate uptake and/or growth are being inhibited by high cell density, which can be due to poorer mass transfer of substrates to the cells due to their close proximity; a phenomenon observed in literature [5]. Therefore, three time-varying parameters were introduced: two capture the respective substrate uptake inhibition of glucose and nitrate ($\theta_G(\cdot)$ and $\theta_N(\cdot)$, respectively); and the other captures biomass growth inhibition ($\theta_X(\cdot)$). These time-varying parameters

are functions of state variables in the form of an ANN. Equations (1), (2), and (3) are therefore rewritten as in

$$\frac{dX}{dt} = \theta_X(\mathbf{S}) \cdot \mu \cdot X - d_X \cdot X - \frac{F}{V} \cdot X \quad (9)$$

$$\frac{dG}{dt} = \left(G_f - G\right) \cdot \frac{F}{V} - \theta_G(\mathbf{S}) \cdot v_{\max_G} \cdot \frac{G}{K_G + G} \cdot X \quad (10)$$

$$\frac{dN}{dt} = \left(N_f - N\right) \cdot \frac{F}{V} - \theta_N(\mathbf{S}) \cdot v_{\max_N} \cdot \frac{N}{K_N + N} \cdot X \quad (11)$$

$$\mathbf{S} = [X, N, G] \quad (12)$$

where G_f and N_f are the feed concentrations of glucose and urea, respectively, F and V are the feed rate and reactor volume, respectively, $\mathbf{S}$ is the vector of state variables, and the remaining parameters in Equations (4)-(5) and (9)-(11) take the same value as those previously identified. Constant volume is assumed. During the identification of time-varying parameters, their deviations are penalised to try and follow the kinetic model dynamics as closely as possible, thus maximizing knowledge transfer, Penalization is done using a penalty term, that is added to the objective function shown in Equation (8), as in

$$\min \frac{1}{n} \cdot \sum_{t,\text{meas}} \sum_i \left(\frac{C_{i_t} - C_{i_{\text{meas}_t}}}{\sigma_i}\right)^2 + \rho \cdot \sum_{t,\text{TVP}} \sum_i \left(\theta_i(t) - \theta_i(t-1)\right)^2 \quad (13)$$

where ρ is a hyperparameter to be tuned. The time-varying parameters are then reformulated as a function of state variables using an ANN. The ANN is embedded within the system of ODEs as illustrated in Figure 1, where state variables update the ANN which updates the time-varying parameters, which can be used in the system of ODEs to update the state variables, thus generating state variable profiles.

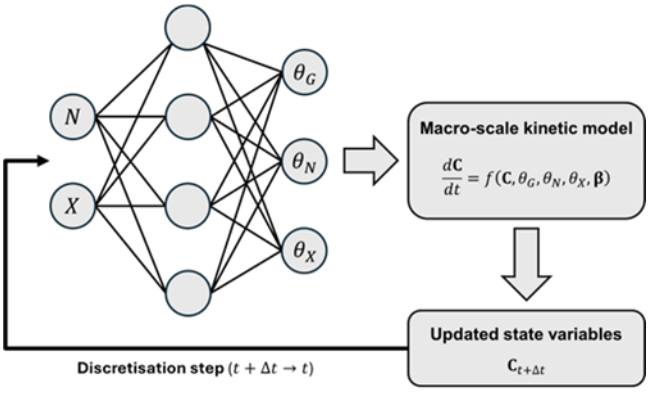

Figure 1. Hybrid model structure: Simulation of time-varying parameters based on system components using an ANN.

Comparing the performance of several ANN architectures allowed the identification of the most appropriate ANN structure. The Akaike Information Criterion with correction for small sample sizes (AICc) was employed to quantify the performance of each ANN structure, as in

$$AICc = n \ln Z + 2\,k + \frac{2\,k^2 + 2\,k}{n - k - 1} \quad (14)$$

where k is the number of parameters with a correction for the small sample size n. The ANN architecture with the lowest AICc score was chosen as the best trade-off between minimal overfitting and accuracy.

2.4 Uncertainty analysis

Uncertainty analysis was conducted using bootstrapping which involved taking samples of the training dataset and training the model on each sample separately, generating a distribution of model outputs, as illustrated in Figure 2. A 95% confidence interval (z_{CI}) of each component z was generated by sampling from the resulting output distribution, as in

$$z_{CI} = \bar{z} \pm c \cdot \frac{\sigma_z}{\sqrt{\frac{n_z - 1}{n_z}}} \quad (15)$$

where $\bar{z}$ is the mean of the component z, c is the confidence level, n_z is the number of samples of the component z, and σ_z is the standard deviation of the component z. In this study, bootstrapping was conducted by removing 48-hour segments of data, a third of the fed-batch time horizon.

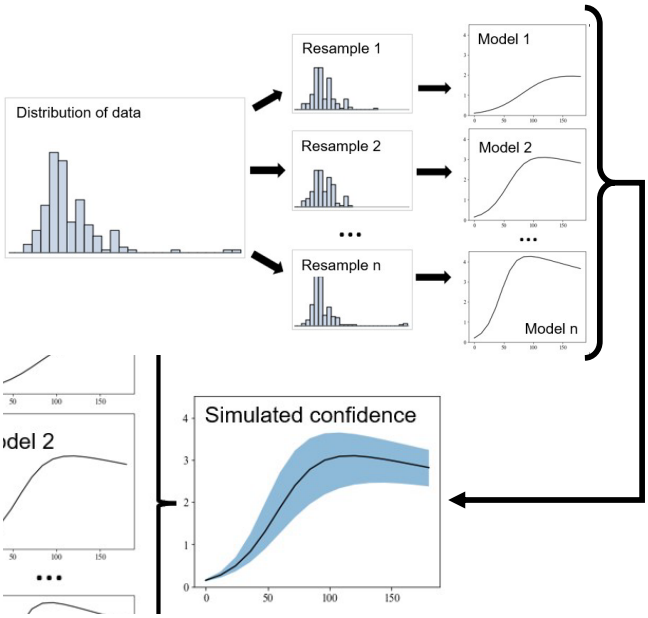

Figure 2. Bootstrapping: Schematic.

2.5 Feed optimization

The dynamic optimization problem follows the general NLP problem described by

$$\min_{\mathbf{F},\mathbf{N_f}} -L(t_f) \quad (15)$$

$$\text{s.t.} \quad 0 \le F(t) \le 50 \text{ mL h}^{-1} \quad (16)$$

$$0 \le N_f(t) \le 60 \text{ g L}^{-1} \quad (17)$$

where $\mathbf{F}$ and $\mathbf{N_f}$ are the vectors of discretized dynamic feed rate and urea feed concentration, respectively, and t_f is the predetermined final time of the batch. Both feed rate and urea feed concentration were allowed to vary every 12 hours, giving the optimization problem 28 degrees of freedom. The dynamic optimization problem was solved using a gradient descent method in the Python optimization package Pyomo. Glucose concentration was also constrained between 3 and 7.5 g L^{-1} beyond a time of 30 h to ensure stable fed-batch operation.

2.6 Methodology summary

The overall modelling methodology was conducted in a manner that requires minimal data collection for model construction and is illustrated in Figure 3.

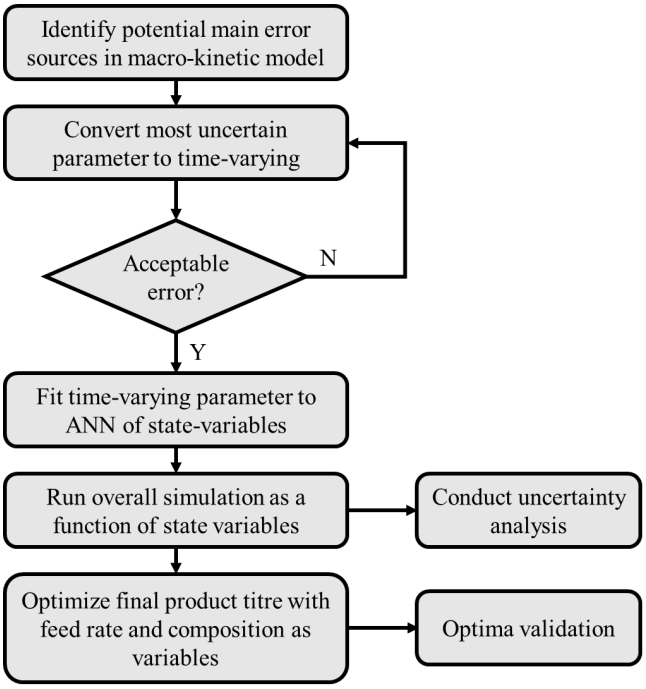

Figure 3. Methodology: Summary of dynamic optimization from hybrid model construction for fed-batch lutein production under uncertainty.

3 RESULTS AND DISCUSISON

3.1 Results of hybrid modelling

The purely mechanistic macro-scale kinetic model was first fit with Equations (1)-(5) with no time-varying parameters. The fitting of the preliminary kinetic model has a mean R^2 value of 0.995 and is shown in Figure 4. Parameter estimation converged after 50 iterations using a derivative-free Particle Swarm Optimization (PSO) algorithm. The selected algorithm has strong exploratory capabilities for parameter estimation in the stiff fed-batch system simulated in this study. The problem was run 5 times, each with 100 particles, to confirm the same optima was being found, increasing the confidence in it being a global optimal solution.

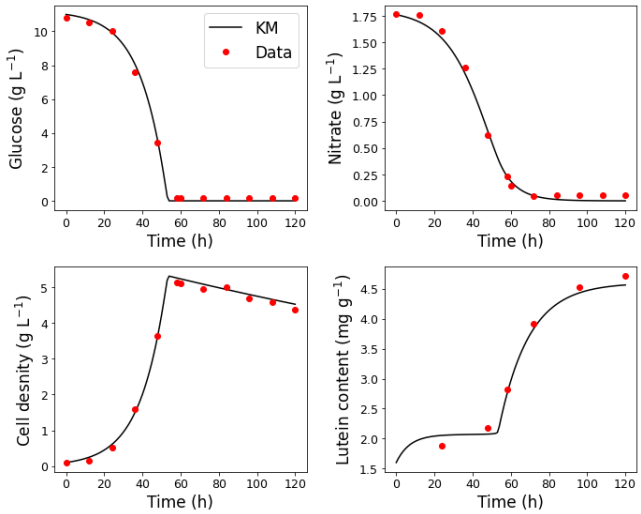

Figure 4. Kinetic model (KM): Plots of substrates glucose and nitrate, biomass and lutein content fitted to the original batch using $NaNO_3$ as a nitrogen source.

The application of this kinetic model to a 5 L fed-batch process, where the cell density reaches levels in excess of 10 times higher than the batch process, illustrated significant process-model mismatch, particularly in the latter stages at the highest cell densities, which is evident in the resulting mean R^2 value of -0.026 for the overall system. This motivated the introduction of time-varying parameters to better fit the more challenging dynamics seen in the latter stages of the fed-batch process. Time-varying parameters were first introduced to a fed-batch operation that reinstated the same initial substrate concentrations as the preliminary batch process after each injection of concentrated medium. Time-varying parameters were initially simulated as a discretized function of time. Parameter estimation converged within 100 iterations of the IPOPT algorithm in the Python package SciPy.

The resulting time-varying parameter profiles were then used to train the ANN. The most appropriate ANN architecture had seven hidden nodes and sigmoid activation functions. The ANN was trained over 1700 epochs using data augmentation, following the method described in a previous study [6]. The fitting of the hybrid model is shown in Figure 5, with the profiles of the corresponding time-varying parameters shown in Figure 6. The introduction of time-varying parameters increased the mean R^2 value for cell density and lutein content from

-0.026 to 0.855. The decreasing trend of the time-varying parameters simulated by the ANN for glucose and biomass highlights the inhibitory effects that the high-cell-density has on substrate uptake and growth. It is observable that all parameters saw an overall decrease in value from start to finish, further emphasizing the inhibitory effects of a HCDC system.

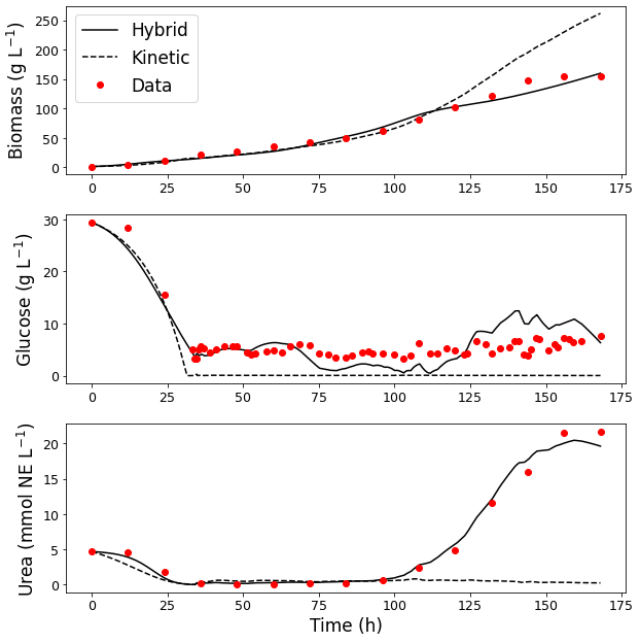

Figure 5. Hybrid model (HM): Plots of substrates glucose and nitrate, and biomass for fed-batch simulation with comparison to the original Kinetic Model (KM). Note that Nitrate Equivalent (NE) refers to the equivalent concentration of nitrogen in a solution of $NaNO_3$.

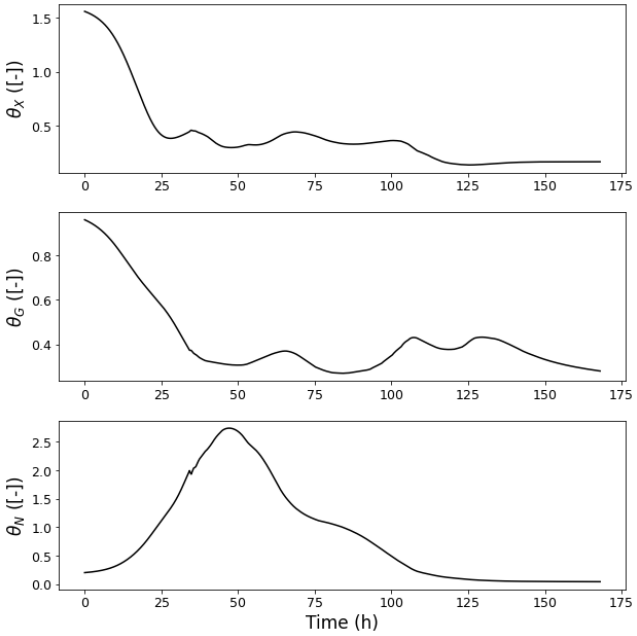

Figure 6. Hybrid model: Time-varying parameters.

3.2 Results of feed optimization

The optimization of feed rate and composition was successfully conducted, converging within 500 iterations of the IPOPT algorithm on the Python package Pyomo. Figure 7 shows the optimized trajectories in comparison to the original experimental conditions. All constraints were successfully satisfied.

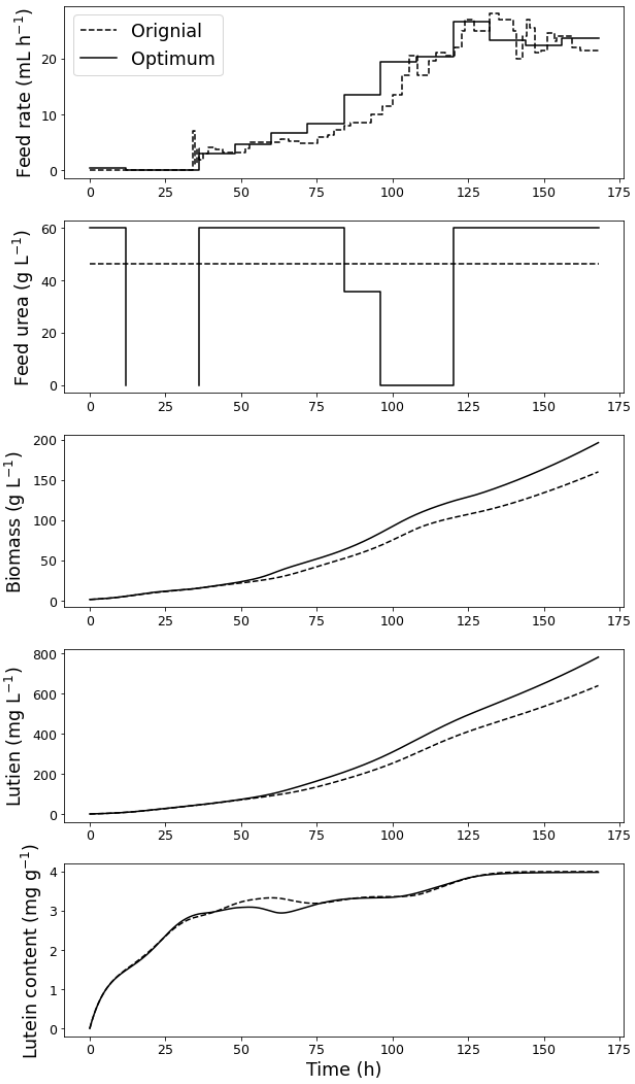

Figure 7. Optimized feeding: Plots of feeding rate and feed urea concentration changing every 12 hours, as well as plots of biomass and lutein concentration and content for both the original and optimized cases. It should be noted that there is no urea feed concentraiton between times 12 h and 36 h since there is no feeding during this period.

Figure 7 shows an increase in lutein production from the original experiment of over 25%. Since the lutein content is almost unchanged, the increased production is coming exclusively from increased biomass growth, which is most noticeable beyond a time of 50 h. This

implies that for optimizing the lutein content of the cell, strain modification would be a more attractive option as external control variables have less impact on the internal lutein content than the biomass growth.

A significant difference between the substrate trajectories can be seen for urea, for which the optimized case demonstrates the maintenance of a low urea concentration within the vessel, while still feeding enough to sustain the biomass. This implies that nitrate may be inhibitory towards optimal product formation, despite being required for biomass growth.

In terms of the feed trajectories, the most noticeable difference between the original and the optimized case for feed rate is between 60-110 h, where the optimized trajectory shows higher flowrate. This causes an increase in growth during this period without causing a spike in glucose concentration within the reactor, thus allowing more lutein to be produced. Urea feed concentration generally lies at the upper bound of 60 g L^{-1}, with the exception of 12-36 h where there is no flow, and 84-120 h where the spike in urea concentration occurred in the original experiment, suggesting that such an increase in urea concentration may be inhibitory to the system.

Although increasing the number of discretization steps for the feed profile could further improve the optima, it would need to be coupled with uncertainty to assess the attainability of a more specific solution.

4 CONCLUSIONS

In this study, a hybrid model was constructed with transfer learning to simulate a fed-batch HCDC system for lutein production. The introduction of time-varying parameters facilitated the capture of increasingly inhibited system dynamics during the latter stages of fed-batch operation. The varying system dynamics were accurately captured, thus maximizing the model's applicability to simulate fed-batch processes under alternate operating conditions. Time-varying parameter values were also penalized, which maximized knowledge transfer and assisted in reducing model overfitting by avoiding dramatic changes between subsequent time-varying parameter values. The kinetic backbone used for the hybrid model construction incorporated as much physical information as possible, meaning minimal additional data was required to train the ANN for both the original hybrid model and the transfer learning updated hybrid model. The ANN allowed the hybrid model to remain a function of observable state variables while being able to capture nonlinear dynamics that cannot be considered by the kinetic backbone. The successful implementation of dynamic optimization facilitated the prediction of optimal feeding trajectories under plausible operating conditions for experimental validation.

Overall, the optimization shows plausible predictions for experimental validation. The success of this modelling strategy and application to fed-batch optimization paves the way towards the development of an effective digital twin for the modelling of fed-batch systems for maximizing lutein production and yield for future fed-batch HCDC operation. The ability of this framework to incorporate a data-driven model with minimal data gives it substantial potential in applications to novel bioprocesses, where both experimental data and mechanistic understanding are limited.

ACKNOWLEDGEMENTS

The authors would like to acknowledge our experimental collaborators in Fuzhou University and Xiamen University, as well as BBSRC for the funding of this work.

REFERENCES

1. Harrington, R. Growing the bioeconomy: a national strategy to 2030, HM Government (2018)
2. Riesenberg, D., and Guthke, R. High-cell-density cultivation of microorganisms. Applied Microbiology and Biotechnology, volume 51, pages 422-430 (1999)
3. Wu, H., Shuhua, Q., Ruixiong, Y., Qihua, P., Yinghua, L., Chuanyi, Y., Ning, H., Song, H., and Xueping, L. Strategies for high cell density cultivation of Akkermansia muciniphila and its potential metabolism. Journal of Clinical Microbiology, volume 12, issue 1 (2024)
4. Zinnecker, T., Reichl, U., and Genzel, Y. Innovations in cell culture-based influenza vaccine manufacturing – from static cultures to high cell density cultivations. Human Vaccines & Immunotherapeutics, volume 20, article 2373521 (2024)
5. Ruhe, Z.C., Low, D.A., and Hayes, C.S. Bacterial contact-dependent growth inhibition. Trends in Microbiology, volume 21, issue 5, pages 230-237 (2013)
6. Rogers, A.W., Song, Z., Vega-Ramon, F., Jing, K., and Zhang, D. Investigating 'greyness' of hybrid model for bioprocess predictive modelling. Biochemical Engineering journal, volume 190 (2020)