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

Hybrid model development for Succinic Acid fermentation: relevance of ensemble learning for enhancing model prediction

Juan Federico Herrera-Ruiz[a], Javier Fontalvo[a], Oscar Andrés Prado-Rubio[a*]

[a] Departamento de Ingeniería Química, Universidad Nacional de Colombia – 170003 Manizales, Colombia
* Corresponding Author: oaprador@unal.edu.co

ABSTRACT

Sustainable development goals have spurred advancements in bioprocess design, driven by improved process monitoring, data storage, and computational power. High-fidelity models are essential for advanced process system engineering, yet accurate parametric models for bioprocessing remain challenging due to overparameterization, often resulting in poor predictive accuracy. Hybrid modeling, combining parametric and non-parametric methods, offers a promising solution by enhancing accuracy while maintaining interpretability. This study explores hybrid models for succinic acid fermentation by *Escherichia coli*, a critical process for sustainable bio-based chemical production. The research presents a structured exploration of hybrid model architectures and their robustness under varying conditions. Experimental data were preprocessed to remove noise and outliers, and hybrid model structures were developed with differing levels of hybridization (from one to all reaction rates). Kinetic parameters were recalibrated and compared against original values. Machine learning algorithms, including Artificial Neural Networks, Support Vector Machines, and Gaussian Processes, were tested, with tuning strategies applied to original or recalibrated parameters. Due to a considerable variability in individual model performance for validation, an ensemble learning approach was proposed to enhance robustness. Results demonstrate that despite not solving the overparametrization issues, all hybrid models outperform the original parametric model, with the best-performing hybrid model achieving a 52.3% lower RMSE of validation, avoiding overfitting. Ensemble approaches further improved predictions, reducing RMSE by up to 62.3% compared to individual parametric models. This highlights hybrid modeling's potential to enhance bioprocess prediction accuracy, even with limited data, supporting future advancements in bioprocess scale-up, digitalization, and sustainable biorefinery implementations.

Keywords: Modelling, Fermentation, Modelling and Simulations, Machine Learning, Reaction Engineering, Hybrid modelling, Succinic Acid, Kinetics.

1. INTRODUCTION

Succinic Acid is considered as one of the key molecules to pave the way for sustainable bio-based production of chemicals. Succinic acid is used to produce over 30 commercially available products, including food ingredients and additives, commodity chemicals like polymers (PBS), solvents and other organic acids and even pharmaceutical intermediates [1].

The increasing focus on sustainable development goals has spurred significant research into bioprocesses optimization, particularly through technological advancements in process monitoring, data storage, and computational capabilities. These developments, combined with modelling techniques and simulation tools, are driving substantial advances on digitalization of biomanufacturing. In this context, hybrid modelling has emerged as a powerful approach, combining parametric and non-parametric methods to mitigate their individual drawbacks. Parametric kinetic models for succinic acid fermentation tend to be overparameterized and uncertain, leading to poor predictive performance [2,3], presenting prediction

https://doi.org/10.69997/sct.153338

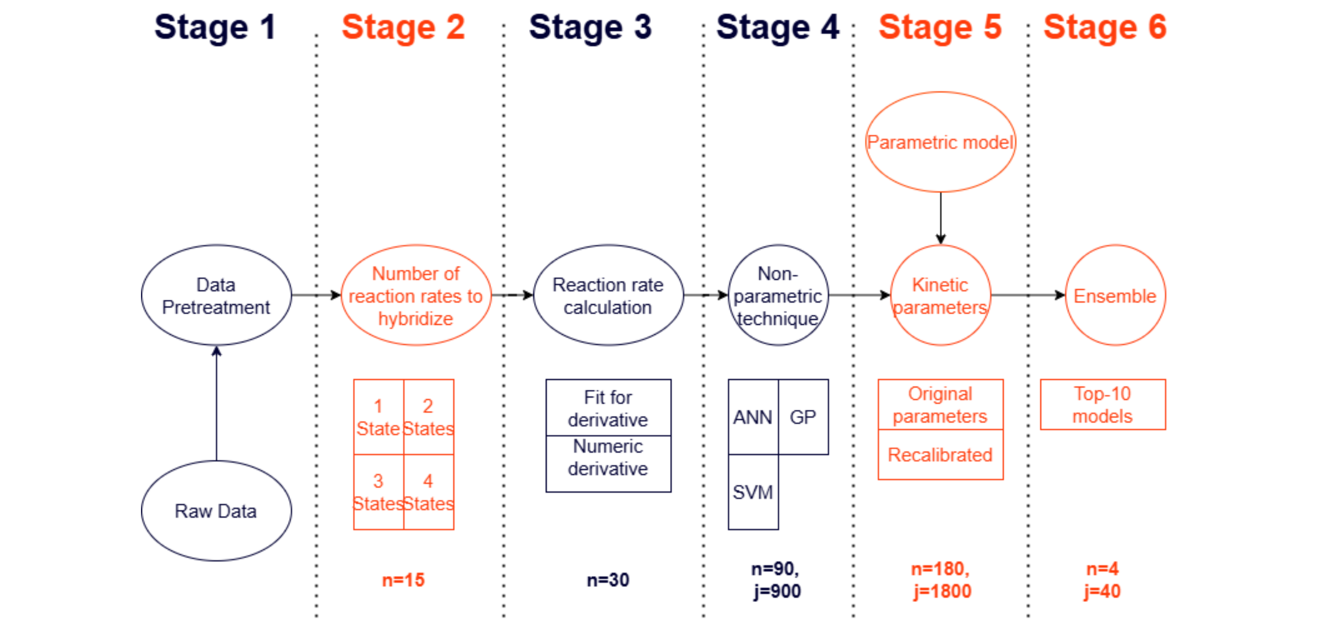

Figure 1: Methodology workflow.

errors up to 30% in key variables such as biomass and product concentrations [4]. Hybrid modelling has been successfully applied to describe fermentations, reducing errors in industrial settings up to 50% when compared against parametric models [5]. Therefore, hybrid modelling rises as an alternative to increase accuracy and precision for describing succinic acid fermentations.

This study focuses on developing a hybrid model to harness limited experimental data and improve states' predictions of succinic acid fermentation by *Escherichia coli* [6]. To achieve this objective, model structures and nonparametric techniques were systematically constructed and investigated. Different levels of hybridization were tested, ranging from 1 to 4 reaction rates replaced by either Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Gaussian Processes (GPs). Ensemble learning was proposed to enhance predictive capabilities, due to high variability between predictions of individual models and poor performance. This paper is structured as follows: section 2 presents the methodology for constructing the hybrid models, detailing data pretreatment and model structure choices. In section 3 results are shown, comparing different model structures and their combinations. Finally, section 4 provides the conclusions of this investigation.

2. METHODOLOGY

Figure 1 presents the workflow proposed in the present work. Stages colored in orange represent structural decisions made to develop the model and stages presented in dark blue represent numeric decisions. Mass balances constitute the parametric knowledge incorporated to the hybrid model whereas reaction rates are replaced by nonparametric techniques. In the lower part of Figure 1, *n* refers to the number of model structures involved in the stage, while j refers to the total number of possible trained models. Notice that the sequential stages from 2 to 5 correspond to model structures building and do not mean those models were trained in each stage. The total number of models trained were 1800 in Stage 5.

2.1 Data Pretreatment

The experimental data was taken from the original paper [6]. First, outliers were detected and removed following established methodologies [7]: i) Outliers were identified using a moving median window of three datapoints and ii) if an outlier was identified, this was replaced using the Savitsky-Golay polynomial approach. Then, data augmentation was performed to enhance the predictive capabilities of the ML algorithms. Originally, experiments ranged from 21 to 37 experimental datapoints and the data augmentation procedure was done, so each experiment had 50 datapoints for each variable. Experiments were separated in training and validation in the same partitions as the original authors [6].

2.2 Reaction rate hybridization

In Stage 2, a decision is made to quantify how many and which reaction rates will be replaced by non-parametric model. There are a total of 15 possible combinations of reaction rates replaced by non-parametric techniques.

As such, the reaction rate of a state is replaced by a non-parametric technique, as exemplified in eq. 1 for Succinic Acid (SA):

$$\frac{dSA}{dt} = f(t, S_{in}) \quad (1)$$

where $f(t, S_{in})$ represents the output of a non-parametric technique that receives time (t) and initial substrate concentration S_{in} as inputs.

2.3 Reaction rate calculation

Reaction rates are used to train the non-parametric techniques chosen in Stage 4. To approximate these rates two different approaches were tested: i) a sigmoidal fit for substrate, succinic and acetic acid, and a polynomial fit for biomass were implemented, and then derived to obtain the reaction rate, and ii) the numeric derivative was calculated using the smoothed experimental data. This procedure yields 30 different model combinations to evaluate.

2.4 Non-parametric technique

In Stage 4, the non-parametric technique used to describe the reactions rates replaced in Stage 2 and calculated in Stage 3 is chosen. In this work, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Gaussian Processes (GP) were tested for each hybrid model. Notice that the same non-parametric technique was used for all the reaction rates, not considering mixtures such as one reaction rate modeled using GPs and other using ANNs. Therefore, there are 90 model combinations to evaluate. Each combination was trained 100 times, yielding 900 different trained models. Hyper-parameters were optimized using MATLAB 2024b® built-in functionalities.

Hybrid Model structure

2.1.1. Parametric model

For the succinic acid fermentation, the original researchers proposed a model to describe biomass growth, succinic and acetic acid production, and substrate consumption composed by 4 ODE's and 16 parameters. The biomass growth is expressed using a modified Monod expression (eq. 2) incorporating the product and substrate inhibition terms proposed by Levenspiel [8]. Cell lysis was considered, and it's expressed in eq. 3:

$$\mu = \mu_{max} \cdot \left(\frac{S}{S+k_S}\right) \cdot \left(1 - \frac{SA}{SA_{crit}}\right)^p \cdot \left(1 - \frac{S}{S_{crit}}\right)^n \quad (2)$$

$$k_L = k_{l,max} \cdot \left(\frac{k_l}{S+k_l}\right) \quad (3)$$

$$\frac{dX}{dt} = (\mu - k_L) \cdot X \quad (4)$$

where, S the substrate concentration, SA the succinic acid concentration, X represents the biomass concentration, μ is the specific growth rate, μ_{max} the maximum specific growth rate, SA_{crit} the critical succinic acid concentration for inhibition, p the power constant of succinic acid inhibition, S_{crit} the critical substrate concentration for inhibition, n the power constant of substrate inhibition..

Product formation of Succinic Acid and Acetic Acid was described using Luedeking-Pirett expressions, shown in eqs 5 and 6, where AA represents the acetic acid concentration, α_{SA} and α_{SA} represent the growth associated constants for AA and SA respectively, whereas β_{AA} and β_{SA} represent the non-growth associated constants for AA and SA, respectively.

$$\frac{dSA}{dt} = \alpha_{SA} \cdot \frac{dX}{dt} + \beta_{SA} \cdot X \quad (5)$$

$$\frac{dAA}{dt} = \alpha_{AA} \cdot \frac{dX}{dt} + \beta_{AA} \cdot X \quad (6)$$

Substrate consumption depends on two terms: one associated with biomass growth (eq. 7) and a term associated with biomass maintenance (eq 8). The substrate consumption rate is shown in eq. 9:

$$w = \frac{1}{Y_{X/S}} + \frac{\alpha_{SA}}{Y_{SA/S}} + \frac{\alpha_{AA}}{Y_{AA/S}} \quad (7)$$

$$z = \frac{\beta_{SA}}{Y_{SA/S}} + \frac{\beta_{AA}}{Y_{AA/S}} + m_{e,X} \quad (8)$$

$$\frac{dS}{dt} = -(w \cdot dX + z \cdot X) \quad (9)$$

where $Y_{X/S}$,$Y_{SA/S}$,$Y_{AA/S}$ represent the yields for biomass, succinic acid and acetic acid respectively. $m_{e,X}$ represents the cell maintenance coefficient of substrate consumption.

2.1.2. Remaining Kinetic Parameters values

For the hybrid model consolidation, the kinetic parameters that remain in the mass balances, after replacing the reaction rates predicted by the ML technique in Stage 4, are evaluated in two different schemes:

- Keeping the original values proposed by the authors.
- Recalibrating the kinetic parameters.

This final procedure yields 180 model combinations, and 1800 models trained.

2.1.3 Ensemble learning approach

In preliminary simulations, high variability in predicted outputs and poor performance of individual models were observed. To address this issue and improve model performance, an ensemble approach was implemented by algebraically averaging predictions. Specifically, 100 models were trained for each model structure, and the top 10 models were selected and evaluated based on their performance. From these, the best combinations were identified for cases involving the replacement of 1, 2, 3, and 4 reaction rates. This process resulted in the selection of 4 model structures and a total of 40 trained models, which were ultimately used for predictions.

2.1.4 Computational Aspects

Simulations were conducted on a commercial laptop with an Intel(R) Core (TM) i7-10510U CPU 1.80GHz, 2304 Mhz with 4 cores, 8 GB of RAM and no dedicated GPU. Calculations were performed in Matlab 2024a®. Models were trained in parallel with training times ranging from 9 up to 49 hours.

To compare the results of the different models, the RMSE reduction was calculated, as seen in eq 10:

$$RMSE_{reduction} = \frac{RMSE_{Parametric\ Model} - RMSE_{Hybrid\ Model}}{RMSE_{Parametric\ Model}} \quad (10)$$

3. RESULTS

Figure 2 shows the SA concentration in a validation experiment for the parametric model, when only the SA reaction rate is replaced using GPs.

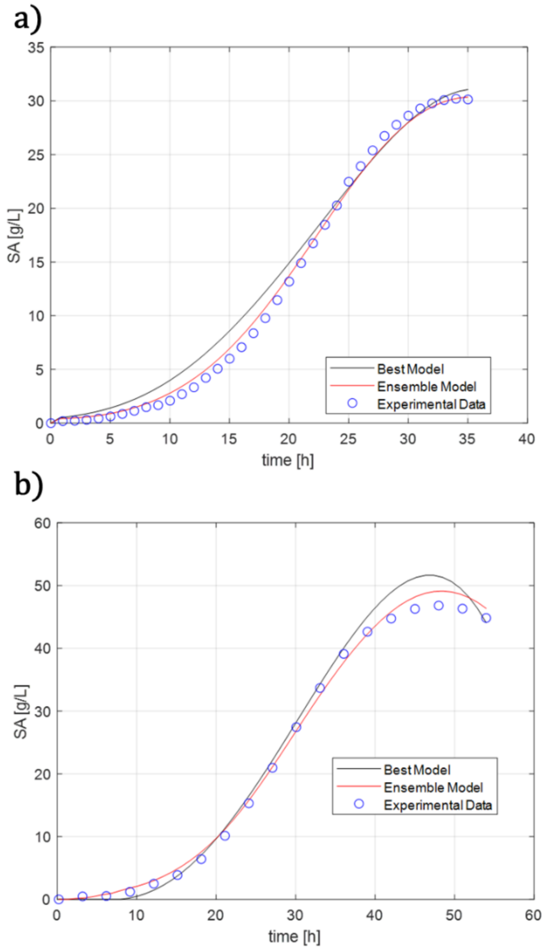

Figure 2. Best individual model fit and ensemble predictor when replacing only the SA reaction rate . a) initial concentration of 40g/l of Xylose. b) concentration of 60g/l of Xylose.

As seen in Figure 2, the best individual model (GPs, numeric derivatives for reaction rates and no kinetic parameter recalibration) correctly predicts the general trends and values of experimental data, but the ensemble model (GPs, numeric derivatives for reaction rates and no kinetic parameter recalibration) is closer to them. Consequently, the parametric model has an RMSE of 7.03, while the best individual fit has an RMSE of 6.68 and the ensemble prediction presents an RMSE of 6.37. Although both the best individual model and the ensemble model perform better than the parametric one, the ensemble approach yields a significantly better RMSE, with an RMSE reduction of 9.9% against the 4.9% yielded by the best individual fit. Overall, ensemble approaches show prediction errors in the validation set ranging from 5% to 14% lower when compared to the best individual models, with the greatest gains observed when replacing all four reaction rates.

Figure 3 shows the ensemble model and its confidence intervals when using GPs and replacing only the SA reaction rate. Confidence intervals are built using the lowest and highest prediction of SA within the ensembled models for each sample time. The median confidence interval presents a 10% deviation from the ensemble model's prediction. Higher deviations are seen for lower times, whereas the model's confidence intervals become narrower as the time advances, becoming as narrow as 2.3%. As a result, 94% of the experimental datapoints lie inside the confidence interval with all the points outside of it being in fermentation times less than 4 hours, signaling that the ensemble model is robust enough to average the weaknesses and strengths of individual models while accurately describing the dataset.

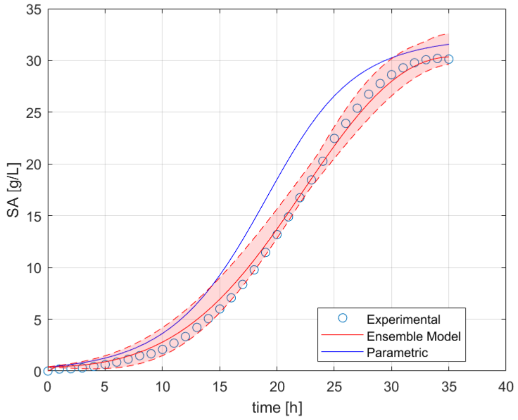

Figure 3. Ensemble predictor uncertainty when replacing only the SA reaction rate

Figure 4 presents the profiles obtained for the ensemble models when replacing SA, and X for ANNs and using numeric derivatives in validation experiments. It is noteworthy that despite the similarities in the profiles, the Hybrid Models have different performances, with an

RMSE of 5.68 (RMSE reduction of 27.1%) for the hybrid model with the original parameters and 5.22 (RMSE reduction of 34.8%) for the hybrid model recalibrating the other initial parameters.

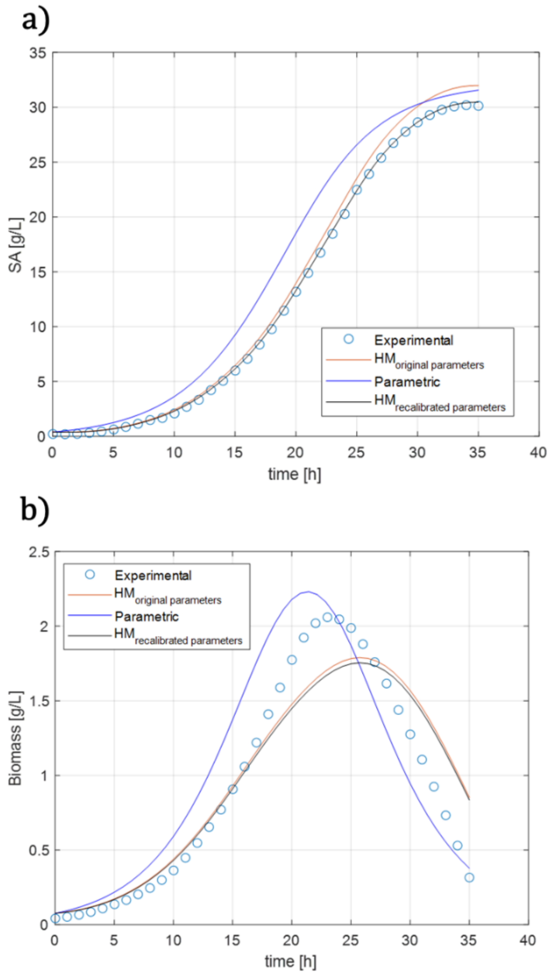

Figure 4. Performance of the ensemble models when replacing SA, and X for ANNs and using numeric derivatives. a) Succinic Acid for an initial concentration of 40 g/L of Xylose. b) Biomass for an initial concentration of 60 g/L of Xylose. HM stands for "Hybrid Model".

As seen in Figure 4, the Hybrid models represent more accurately the behavior of SA, whereas the parametric model tends to overpredict concentrations. Nonetheless, the Hybrid Model that uses the original kinetic parameters results in higher final SA concentrations than both the parametric model and the experimental data. For biomass, the parametric model tends to overestimate concentrations and predicts the biomass peak earlier than the experimental data shows, with a steep descent that results in an underestimation of the biomass concentrations. Hybrid Models perform better up until the 20h mark, where the Hybrid Models deviate, predicting lower concentrations and a lower peak displaced around 5h. After the predicted peak, the Hybrid Models tend to predict higher concentrations than both the parametric model and experimental data.

When recalibrating the kinetic parameters, their values differed on average 14.9% when compared against the original ones. The mass balances show high dependency between one another, therefore replacing reaction rates could provide better description for a specific variable while worsening other predictions.

Table 1 shows the RMSE reduction in validation classified by number of reaction rates and machine learning technique for the best ensemble models. The best ensemble models for the three machine learning techniques are as follows, with all of them using numeric derivatives and recalibrating kinetic parameters:

- Replacing only one reaction rate (SA).
- Replacing two reaction rates (SA and AA)
- Replacing three reaction rates (SA, AA and X)

Table 1: Percentage of RMSE reduction in validation experiments classified by number of reaction rates replaced and nonparametric technique used.

Black Box Technique	1 rate	2 rates	3 rates	4 rates
ANN	20.9%	34.8%	61.4%	62.3%
GP	15.6%	29.8%	50.9%	57.0%
SVM	13.3%	27.1%	40.5%	47.3%

As expected, for all machine learning techniques, the model predictions improve by increasing the number of reaction rates replaced. However, there are diminishing returns each time, since the biggest performance gains are seen when jumping from 2 to 3 reaction rates, reducing RMSE around 50% for all cases. This is due to the states interact between each other in the model, therefore, a better description of one state does not necessarily translate into better description of the other states. This occurs despite recalibrating the remaining model parameters, as can be seen in Figure 3 for biomass.

Generally, ANNs perform better than SVMs and GPs for all cases. However, the performance of ANNs and GPs is comparable, with latter needing fewer intensive calculations. When replacing the four reaction rates, GPs reduce the RMSE 57%, with ANNs presenting a 62.8% reduction despite taking almost 5 times larger training time (9 hours versus 45 hours). SVMs have comparable performance to the other two machine learning techniques up to replacing two states, however the performance gains when using three or more are less than what GPs and ANNs present.

Despite gaining better performance when replacing more reaction rates, the structure of the model becomes less parametric and approaches more to a full non-parametric technique. This means that the model loses interpretability and explainability in exchange for more numeric precision. Moreover, the performance gains seen from three to four reaction rates replaced are small, signalling diminishing returns.

4. CONCLUSIONS

The results of this study underscore the transformative potential of hybrid modeling in succinic acid fermentation by *Escherichia coli*. The hybrid models demonstrated superior predictive accuracy, achieving RMSE reductions of up to 62.3% when replacing four reaction rates, significantly outperforming traditional parametric models. Also, the individual hybrid models suffer from low robustness during validation, making necessary the use of ensemble learning, which is not commonly use in ML applications. Ensemble approaches provided higher accuracy, by leveraging the strengths of the individual models while minimizing their weaknesses. Ensemble approaches could be very useful for developing bioprocesses because they are usually less sensitive to noisy data, making them more robust than traditional models. Among the non-parametric techniques tested, Artificial Neural Networks (ANNs) delivered the highest accuracy but at a greater computational cost, whereas Gaussian Processes (GPs) provided comparable performance with reduced complexity, making them an efficient alternative. The recalibration of kinetic parameters emerged as a critical factor for optimizing model performance, particularly when paired with numeric derivatives for reaction rate calculation. While increasing the number of replaced reaction rates improved numerical precision, this came at the expense of model interpretability, with diminishing returns observed beyond three reaction rates. Besides using black box techniques to replace reaction rates does not solve the issue of overparametrization of kinetic models in bioprocesses which might lead to overfitting and poor prediction capabilities in the validation sets. However, it has been shown that using ensemble approaches helps to reduce models' individual weaknesses and enhance their strengths, therefore increasing and enhancing prediction capabilities, as seen by better predictions in the validation set with reduced predictor uncertainty.

ACKNOWLEDGEMENTS

This work was funded by the Sistema General de Regalías de Colombia (SGR) fund, through the Project BPIN: 2022000100035.

REFERENCES

1. Dai Z, Guo F, Zhang S, Zhang W, Yang Q, Dong W, et al. Bio-based succinic acid: an overview of strain development, substrate utilization, and downstream purification. Biofuels, Bioproducts and Biorefining 2020;14:965–85. https://doi.org/10.1002/bbb.2063.
2. de Azevedo CR, Díaz VG, Prado-Rubio OA, Willis MJ, Préat V, Oliveira R, et al. Hybrid Semiparametric Modeling: A Modular Process Systems Engineering Approach for the Integration of Available Knowledge Sources. Systems Engineering in the Fourth Industrial Revolution, Wiley; 2019, p. 345–73. https://doi.org/10.1002/9781119513957.ch14.
3. Leonov P. Bio-succinic acid production from alternative feedstock. Denmark Technical University, 2022.
4. F. Vigato, J.M. Woodley, M. Alvarado-Morales, Modeling the effect of CO2 limitation in continuous fermentation for biosuccinic acid production, Journal of CO2 Utilization 79 (2024) 102651. https://doi.org/10.1016/j.jcou.2023.102651.
5. P. Shah, M.Z. Sheriff, M.S.F. Bangi, C. Kravaris, J.S. Il Kwon, C. Botre, J. Hirota, Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters, Chemical Engineering Journal 441 (2022). https://doi.org/10.1016/j.cej.2022.135643
6. Chaleewong T, Khunnonkwao P, Puchongkawarin C, Jantama K. Kinetic modeling of succinate production from glucose and xylose by metabolically engineered Escherichia coli KJ12201. Biochem Eng J 2022;185:108487. https://doi.org/10.1016/j.bej.2022.108487.
7. Sánchez-Rendón JC, Morales-Rodriguez R, Matallana-Pérez LG, Prado-Rubio OA. Assessing Parameter Relative Importance in Bioprocesses Mathematical Models through Dynamic Sensitivity Analysis, 2020, p. 1711–6. https://doi.org/10.1016/B978-0-12-823377-1.50286-X.
8. Levenspiel O. The monod equation: A revisit and a generalization to product inhibition situations. Biotechnol Bioeng 1980;22:1671–87. https://doi.org/10.1002/bit.260220810.