Hybrid machine learning for scale-up of biomass production using photobioreactors

Abhishek Sivarama , Alireza Mehrdadfara , Lara Grecoa, Christian Eulerb, Seyed Soheil Mansouria,\*

aDepartment of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads, Building 228A, 2800 Kgs. Lyngby, Denmark

bDepartment of Chemical Engineering, University of Waterloo, 200 University Avenue West Waterloo, ON, Canada N2L 3G1

\*seso@kt.dtu.dk

Abstract

Artificial intelligence (AI) and machine learning (ML) have found widespread acceptance in the field of chemical engineering for the scale-up of processes. However, the translation of these scale-up strategies to biological processes has been challenging due to the emergence of unexpected phenomenology, including new metabolic pathways, during scale-up. While a key issue is the lack of interpretability and explainability in developed algorithms, the primary challenge lies in the transferability of ML models across different scales. To make these models readily deployable, it is crucial to incorporate comprehensive process information while harnessing the potential of AI. In this work, we present insights into a framework which uses hybrid models combining first-principles knowledge about photobioreactors and AI, validated using experiments at different process scales.

We present this framework in the context of scale-up of biomass production in photobioreactors. Using small-scale experiments with volumes of 500 mL and 3 L; a medium-scale experiment with 30 L experiment is also conducted, which is used to validate our approach. This comprehensive approach not only addresses the challenges of biological process scale-up but also ensures the reliability and adaptability of the developed models for real-world applications.

**Keywords**: *Artificial Intelligence, Hybrid modeling, Process scale-up, Photobioreactors*

* 1. Introduction
		1. Scale-up of processes and artificial intelligence

AI has gained a lot of popularity in quite a few use-cases in biotechnology and bioprocess engineering. The use of AI algorithms is usually driven by specific use-cases/problems of interest, resulting in a significant lack of transferability of results from the various algorithms across scales of operation. This problem is in addition to the lack of sufficient data across scales to train major machine learning models. Dynamical modeling, however, has the potential to minimize the requirement of data for modeling. This is because along with fitting the experimental data, dynamical models must be able to fit the dynamics of the system as well. This additional constraint limits the possible trajectories the system can take (Cuomo et al., 2022; Jin et al., 2021). Another issue with using off-the-shelf algorithms for scale-up is the lack of explainability of why certain solutions are achieved. Various recent works aim to increase global explainability of model forms(Kaiser et al., 2018; Sivaram & Venkatasubramanian, 2022), specifically toward the estimation of process dynamics. These dynamical models aid in the development of efficient control strategies and their reduced order nature also help in identifying fast solutions.

* + 1. Photobioreactors as a case-study

A photobioreactor is a high-tech algae incubator. It is a system that uses light to cultivate and grow microorganisms, such as algae or bacteria, to obtain valuable chemical products. The light provides energy for photosynthesis, allowing these microorganisms to thrive and multiply. These photobioreactors have been used in fields like biofuel production, wastewater treatment, and even in some experimental setups for studying and optimizing the growth of microorganisms. Typically, biomass growth models are written in terms of the specific growth rate of the biomass $μ$,

|  |  |
| --- | --- |
| $\frac{dx}{dt}$ $=μx$ | (1) |

where $x$ is the amount of biomass. The specific growth rate $μ$ (h-1) is a function of the amount of substrate, the amount of intensity of light, etc. Depending on the physics of the process, the biomass growth can be described by kinetics like Monod, Michelis-Menten, Andrews, Droop, Caperon and Meyer, Flynn, Martinez kinetics, etc. (Lee et al., 2015). In these models, specific growth rate are functions of substrate concentrations. In photobioreactors, these models are modified to include the relationship of the specific growth rate on the intensity of the incident light. Using these kinetics and including the transport of the substrates, the models for a photobioreactor take the form of ordinary/partial differential equations that include input and output of biomass, substrates (used for the growth of biomass), and constitutive equations for rate parameters. However, these models are not amenable to tracking biomass concentrations due to observability limitations of the models. In the subsequent section, we show how Eq. (1) can be used across scales, in the process eliminating the observability limitations.

* 1. Dataset and Methodology
		1. Dataset description

The literature dataset for this article was collected from (Blanken et al., 2016). The data comes from sources with different operating conditions, operating modes, and process geometry, among other variations across experiments. These include datasets: **70 RPM data, 110 RPM data, 140 RPM exp data, pond M8 data, Pond waste data, A data, B data, C data, D data, E data, F data, G data.** These data come from microorganisms like *Chlorella sorokiniana, Chlamydomonas reinhadtii,* and *Phaeodactylum tricornutum,* with geometries of rotating tubular reactor, laboratory-scale pond reactor, torus-shaped reactor, flat panel reactors and bubble columns.

Apart from these experimental data from literature, we also use a separate experiment on our inhouse system of a 30 L bubble column photobioreactor by Synoxis Algae with *Chlorella sorokiniana*. We see that the dataset has considerable variability in operating conditions, choice of the organisms, geometric configurations, and modes of the experiment.

* + 1. Model simplification – ignoring substrate concentrations

In this work, we consider this specific model form shown in Eq. (1) with experimental results across various photobioreactors, with the goal of identifying invariants across different reactors to determine unchanged properties across scales. Identifying these invariants would aid in generating scale-agnostic models. Models of reactors conventionally incorporate phenomenology for reaction, transport phenomena etc. Substrate concentrations are not always measured; however, we can estimate a nonlinear dynamical model for just biomass growth.

A different dynamic is given by logistic growth rate, which in turn with the decay of biomass function with death rate parameter $k\_{d}$, gives rise to the functional form given by the ordinary differential equation, $\frac{dx}{dt}=k\_{b}\left(1-\frac{x}{K}\right)x-k\_{d}x$, where $k\_{b},K$ are constants related to birth rate and steady-state value of the biomass, respectively. It is known that the logistic growth model is a specific case of the Monod/Michelis-Menten kinetics, under the condition of proportional substrate usage (Alvarez-Ramirez et al., 2019; Putz et al., 2007). The logistic curve implicitly incorporates the availability of the substrate in the term $\left(1-x/K\right)$ where the greater the biomass the lower the amount of substrate availability. The specific growth rate in this situation can be written as, $μ=k\_{b}\left(1-x/K\right)-k\_{d}$. However, these dynamics may not encompass all the possibilities of the interaction with higher-order dependence of the specific growth rate on substrate. This can be surpassed by considering the specific growth model can in general be a polynomial function dependent on the biomass concentration. This reduced order representation allows the biomass growth rate to be dependent directly on the biomass concentration.

* + 1. Light Intensity as a decision variable for the process

Various light intensity models have been developed to codify the dependence of the specific growth rate with intensity. An example model is from (Blanken et al., 2016) where the microalgae system grows as per Eq. 1, except the specific growth rate is a function of the intensity of incident light. The intensity is a function of the biomass concentration, as an increase in the concentration results in lower light penetration, hence “light resource” availability for the growth of the biomass. In this context, one can see that the light availability for the dynamics, itself is like substrate availability. While we can parametrize substrate availability in terms of biomass concentration, in our work we consider light availability as an extrinsic input to the bioprocess. The intensity is normalized to give $u$, where $u$ is 0 when no light is given to the system, and 1 when light is given to the system. Any intermediate intensity for growth can also be captured in the gradation of $u$. We consider this situation where light is continuously provided to the system, i.e., $u=1$. With this context, our parametric specific growth rate can be assumed to be a polynomial function dependent on the amount of biomass, $x$ (which in turn is a substitute for substrate availability) and light availability $u,$ i.e., $μ≔f\left(x,u\right)$.

* + - 1. Overall model and estimation of parameters

There are many ways to approximate the functional dependence of specific growth rate on biomass concentration and light availability. A very popular choice in literature for such hybridized models is to incorporate a neural network for this forward function approximation problem (Cuomo et al., 2022; Nielsen et al., 2020). However, this inherently results in reduced explainability and interpretability of the model form. Although linear models have the advantage of increased interpretability, they suffer from low fit to the data. Inspired by earlier work (Kaiser et al., 2018; Sivaram & Venkatasubramanian, 2022), we consider the specific case of using higher-order polynomial features for this approximation. Benefits here are twofold: 1. The model is linear in parameters, hence aids explainability; 2. Polynomial features accommodate nonlinear dynamics and, hence would result in a better fit. We, therefore, use the reduced order model of specific growth given by a second order polynomial, $μ=c\_{0}+c\_{1}x+c\_{2}u+c\_{3}x^{2}+c\_{4}xu+c\_{5}u^{2}$. Note that this could be in general any higher-order polynomial, but we demonstrate the approach using a second order polynomial function.



Figure 1: Model simplification and process of coefficient estimation

The general architecture of the methodology is shown in Figure 1. With the reduced order polynomial representation of the specific growth rate, and the dataset corresponding to different sampling points given by $\{t\_{k},x\_{k},u\_{k}\}\_{k=1}^{N}$ for each experiment, coefficients corresponding to highly correlated features are systematically set to zero (feature selection). For example, as $x$ and $xu$ are highly correlated, we push $c\_{1}$ to zero. The model is *integrated* for each experiment, and the objective function of the mean-squared error of the biomass prediction at the time points and the corresponding biomass values is minimized. It is noted that given fewer sample points, the model could overfit the data. Hence, a regularization penalty of the one-norm of the coefficients (LASSO) is utilized to have a simpler model that could explain the dataset, resulting in finally the coefficient contributions and estimated trajectory for each experimental dataset.

* 1. Results and discussion



Figure 2: (**a**) Biomass model fit (line) using literature data (marked x) and (**b**) the estimated parameters from the reduced order polynomial model (**c**) Estimated specific growth rate $\hat{μ}$ as a function of biomass concentration $x$ and incident light $u$ for different experimental data



Figure 3: Model fit, parameters, and specific growth rate functional dependence from the in-house experiment on 30 L photobioreactor with *Chlorella sorokiniana*

The estimated model works across scales, geometries, and process conditions of different biomasses (Figure 2**a**), and shows model invariance across scales (Figure 2**b**), demonstrated by similar trends and contributions of relevant features. The specific growth rate functional plots (Figure 2**c**) for each of the individual datasets further enhance our claims of the identified invariance, as the form of the plots remains the same across scales. All the identified characteristics of the model form also fit the in-house 30 L photobioreactor (Figure 3). The estimated steady state value of the biomass concentration was also observed in the 30 L photobioreactor, showing the validity of the estimated model.

This reduced order model for the invariance of the specific growth rate can be written as $μ=α+βu+γxu$. Reasonably, it is seen that the incidence light term and the multiplicative term of the biomass concentration and incident light contribute the most to our predictions. The functional form can be rewritten as $μ=k\_{b}\left(1-x/K\right)u-k\_{d}$, suggesting logistic dynamics with a multiplicative factor based on the incident light. This functional form is also mechanistically viable in that when there is no incident light, i.e., $u=0$, the biomass is bound to decay and not grow, an insight that is driven purely from the data-driven approach. In contrast, this constraint must be manually built into mechanistic models. Depending on the estimated parameter values of the dynamics, we see that the specific growth rate has a nonlinear relationship due to the presence of the $xu$ term in the model form. The degree of nonlinearity is dependent on the magnitude of this term. This model form specifically allows us to know when the system reaches a steady state. For this, the specific growth rate should be zero. We see that such a model form would reach steady-state when $x^{\*}=-\frac{1}{γ}\left(\frac{α}{u^{\*}}+β\right)=\left(1-\frac{k\_{d}}{k\_{b}u\*}\right)K$, where $u^{\*}$ is the steady state light operating input. The maximum value of the steady state is obtained when light is continuously on, i.e., $u^{\*}=1$.

* 1. Conclusions

We have obtained a readily transferable, scale-agnostic, and mechanistically plausible biomass growth model, from experimental data across different scales of photobioreactors. The estimated model form maps directly to a logistic growth model, including a dependence on incident light and death rate. The next step based on the invariance of the model form is the development of online control schemes based on different models fit during experimentation. Different model parameters (dependent on scale) would also help in understanding when the process should be discontinued/stopped for efficient management of resources across different batches of cultivation.

* 1. Acknowledgements

The authors would like to acknowledge that this work was funded in part by Grønt Udviklings- og Demonstrationsprogram (GUDP) under project EXTEND and the European Regional Development Fund (REACT-EU RF-21-0025 “Biosolutions Zealand”).

* 1. References

Alvarez-Ramirez, J., Meraz, M., & Jaime Vernon-Carter, E. (2019). A theoretical derivation of the monod equation with a kinetics sense. *Biochemical Engineering Journal*, *150*. https://doi.org/10.1016/j.bej.2019.107305

Blanken, W., Postma, P. R., de Winter, L., Wijffels, R. H., & Janssen, M. (2016). Predicting microalgae growth. *Algal Research*, *14*, 28–38. https://doi.org/10.1016/j.algal.2015.12.020

Cuomo, S., di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., & Piccialli, F. (2022). *Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What’s next*. http://arxiv.org/abs/2201.05624

Jin, X., Cai, S., Li, H., & Em, G. (2021). NSFnets ( Navier-Stokes flow nets ): Physics-informed neural networks for the incompressible Navier-Stokes equations. *Journal of Computational Physics*, *426*, 109951. https://doi.org/10.1016/j.jcp.2020.109951

Kaiser, E., Kutz, J. N., & Brunton, S. L. (2018). Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *474*(2219). https://doi.org/10.1098/rspa.2018.0335

Lee, E., Jalalizadeh, M., & Zhang, Q. (2015). *Growth kinetic models for microalgae cultivation : A review*. *12*, 497–512. https://doi.org/10.1016/j.algal.2015.10.004

Nielsen, R. F., Nazemzadeh, N., Sillesen, L. W., Andersson, M. P., Gernaey, K. V., & Mansouri, S. S. (2020). Hybrid machine learning assisted modelling framework for particle processes. *Computers and Chemical Engineering*, *140*, 106916. https://doi.org/10.1016/j.compchemeng.2020.106916

Putz, M. V, Lacrămă, A.-M., Ostafe, V., Putz, M. V, Lacrămă, A. M., & Ostafe, V. (2007). Introducing logistic enzyme kinetics. In *Article in Journal of Optoelectronics and Advanced Materials* (Vol. 9, Issue 9). https://www.researchgate.net/publication/228470722

Sivaram, A., & Venkatasubramanian, V. (2022). XAI-MEG: Combining symbolic AI and machine learning to generate first-principles models and causal explanations. *AIChE Journal*, *68*(6). https://doi.org/10.1002/aic.17687