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 machine-learning for dynamic plant-wide biomanufacturing

Shabnam Shahhoseyni[a*], Arijit Chakraborty[b], Mohammad Reza Boskabadi[a], Venkat Venkatasubramanian[b], Seyed Soheil Mansouri[a]

[a] Department of Chemical and Biochemical Engineering, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark
[b] Department of Chemical Engineering, Columbia University, New York, NY 10027, United States of America
* Corresponding Author: shabsh@kt.dtu.dk

ABSTRACT

This study focuses on biomanufacturing case study, i.e. Lovastatin production, employing a hybrid modeling framework that combines mechanistic and data-driven approaches. A time-series dataset was generated using the KT-Biologics I (KTB1) plantwide model, a dynamic simulation of continuous biomanufacturing. The dataset captures critical parameters such as nutrient concentrations and API production. The AI-DARWIN framework was used to develop interpretable machine learning models with constrained functional forms, ensuring both accuracy and clarity. The resulting polynomial-based models reveal key relationships between process variables and system performance, bridging mechanistic insights with data-driven predictions. The models demonstrated reasonable accuracy showing minimal difference between the training and testing errors, highlighting their strong generalization. This work advances hybrid modeling in biomanufacturing by integrating plant-wide mechanistic simulations with interpretable machine learning. The approach ensures both accuracy and transparency while enabling robust process monitoring and control at a 'plant-wide' level, contributing to the broader adoption of hybrid modeling in biomanufacturing.

Keywords: Hybrid modeling, Biomanufacturing, Plant-wide modeling, Lovastatin production, Interpretable machine learning.

1. INTRODUCTION

The biomanufacturing of high-value products, such as Lovastatin, presents unique challenges due to the complex interplay of biological, chemical, and physical processes [1]. Lovastatin, a widely used cholesterol-lowering drug, is produced through microbial fermentation, where variations in nutrient concentrations, reactor conditions, and downstream processing steps significantly influence its yield and quality. Developing accurate and interpretable models to predict and optimize production outcomes is essential for improving efficiency and meeting stringent quality standards.

Traditional modeling approaches in biomanufacturing often rely on mechanistic models, which are grounded in physical and biochemical laws. Developing these mechanistic models can sometimes be challenging, primarily due to the lack of advanced sensing technologies to gather data (for parameter estimation, model validation, etc.) and a limited understanding of multi-scale phenomenological dynamics. However, data-driven modeling has shown great promise in capturing the behavior of complex systems, if and only, there is sufficient amount of data and right measurement of key characteristic phenomena in a process [2]. So, incorporating domain expertise into the modeling framework is crucial.

Bioprocesses, as a prime example of complex systems, are ideal candidates for data-driven modeling, which help bridge gaps in both theoretical knowledge and practical modeling [3]. Although this approach can handle complex, nonlinear relationships in data, It often lacks transparency, making it difficult to trust and interpret its predictions in regulatory and operational contexts. To address this limitation, hybrid modeling approaches that combine mechanistic models with data-

driven techniques have gained significant attention. By combining numeric AI (machine learning) with symbolic AI, a hybrid AI approach can be developed, resulting in robust, interpretable models suitable for complex systems [4].

KT-Biologics I (KTB1) is a dynamic mechanistic simulation model for continuous biomanufacturing, covering the entire Lovastatin production plant [1]. While effective in capturing process mechanisms, KTB1 has long running times and does not fully clarify how process variables influence the final output, limiting its use for real-time monitoring and optimization. To address these limitations, a hybrid modeling approach integrates mechanistic insights with data-driven modeling. The AI-DARWIN framework [5] enables the discovery of interpretable polynomial models by constraining functional forms, ensuring both accuracy and clarity. This approach accelerates computation, enhances understanding of variable relationships, and improves process control efficiency.

This work develops a hybrid model for Lovastatin biomanufacturing, aiming to create an interpretable representation of the process that captures its dynamic behavior, variable interactions, and plant-wide complexities. The approach begins by generating time-series datasets from the KTB1 model under varying nutrient conditions to systematically examine the effects of key variables, such as lactose and adenine, on Lovastatin production. These datasets feed into the AI-DARWIN framework, which builds explainable ML models, laying the foundation for future optimization of nutrient utilization and enabling knowledge transfer to new process scenarios, such as using different microorganisms or producing alternative products.

2. SYSTEM DESCRIPTION

2.1. KTB1 Model Description

This model presents a comprehensive framework encompassing both the upstream production and downstream purification of Lovastatin, a target compound produced via fermentation. As depicted in Figure 1, the upstream process begins with the introduction of key raw materials—adenine and lactose—into a mixing unit (M-101). The mixed medium is then continuously fed into a Continuous Stirred-Tank Reactor (CSTR, labeled R-101), where the biomass, predominantly *Aspergillus terreus*, facilitates Lovastatin production.

The fermentation broth exiting the CSTR is directed to a hydro-cyclone separator (HC-101), which plays a pivotal role in separating the mixture into two streams. The first stream, rich in biomass, is partially recycled back to the CSTR to sustain fermentation. The second stream, with lower biomass content (stream 5), is transferred to the downstream process.

In the downstream phase, the focus is on a sequence of separation and purification steps. Centrifugation units (C-101 and C-102) initially remove remaining biomass and dense impurities from the solution. Following centrifugation, the process advances to nanofiltration (NF-101), where Lovastatin is concentrated by retaining larger molecules through steric exclusion while smaller impurities pass through. Figure 1 illustrates the complete process flow for the biomanufacturing of Lovastatin, including both the upstream and downstream operations.

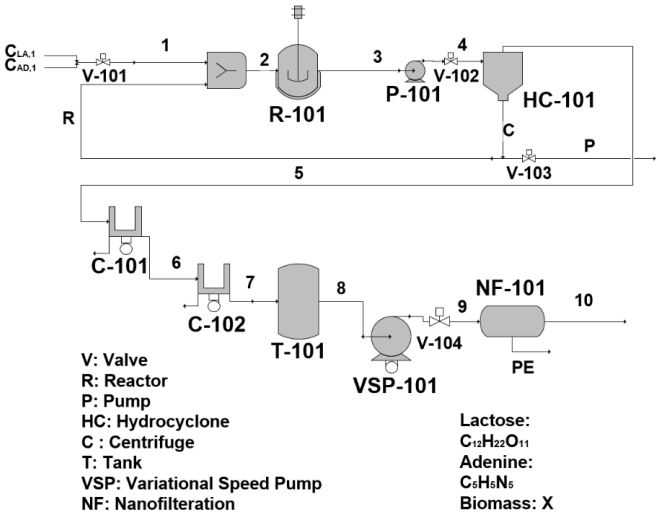

Figure 1. Process flow for the biomanufacturing of Lovastatin

2.1.1 Dataset

The datasets used in this study were generated to systematically explore the relationships between nutrient concentrations and Lovastatin API production. Two distinct datasets were created using the KTB1 mechanistic model, each designed to capture specific aspects of the nutrient dynamics in the biomanufacturing process. The datasets include concentrations of various nutrients (such as lactose and adenine), biomass levels in different streams, and the Lovastatin API produced production plant. In these datasets, only one nutrient concentration was varied at a time (by step change) while keeping all other process parameters constant. This method of isolating single variables ensures a controlled environment to study the individual effects of lactose and adenine on Lovastatin production. Each full dataset consists of about 20000 data points. We used 75% for training and 25% for testing, resulting in 15000 training samples and 5000 test samples. Dataset 1 contains information on altering lactose concentration (CLA). And Dataset 2 contains information on altering adenine concentration (CAD). The graphs corresponding to the generated datasets are presented in Figure 2. A step change in lactose at the start of the process, along with a noisy but constant adenine

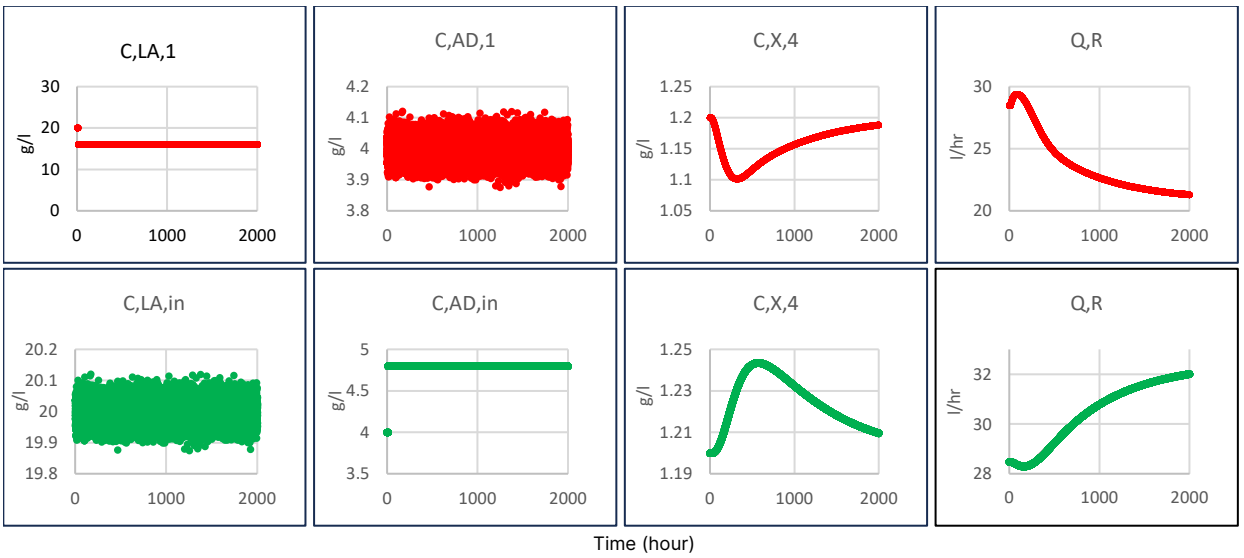

Figure 2. Variables in Dataset 1 (in pink) and Dataset 2 (in greens) were varied, with other variables held constant.

input, were introduced to the model. The noise was intentionally added to simulate real-world scenarios, particularly in nutrient supply, where fluctuations are often observed in one of the inputs. Step changes in lactose and adenine, as seen in the first and second datasets, can arise from sensor malfunctions, control system errors, or other process deviations, which are common faults in industrial processes. These variations result in changes to other system variables, such as biomass concentration and flow rates. Figure 2 illustrates how these variations cause dynamic changes in some of the other variables (C,X,4 and Q,R).

2.2. AI-DARWIN Framework

The AI-DARWIN framework is a hybrid AI mechanistic model discovery tool designed to uncover interpretable and scientifically valid models from data. Unlike traditional black-box approaches, AI-DARWIN focuses on creating explainable models, ensuring that the resulting mathematical representations are not only accurate but also comprehensible to domain experts. This characteristic makes it particularly suitable for hybrid modeling tasks, such as those encountered in biomanufacturing processes. The framework operates by constraining the model discovery process to specific functional forms, such as polynomials. These constraints allow the generated models to remain aligned with physical laws and domain knowledge, fostering trust and usability in critical applications. By imposing such restrictions, AI-DARWIN ensures that the final models capture essential relationships without unnecessary complexity, making them easier to analyze and validate. For the Lovastatin biomanufacturing process, AI-DARWIN is applied to time-series data generated from the KTB1 mechanistic model. The framework systematically evaluates the dataset to uncover explicit mathematical relationships between input variables, such as nutrient concentrations (e.g., lactose and adenine), and output variables, such as Lovastatin API production. Moreover, normalization using standardization (subtracting the mean of the training data and dividing the difference by the standard deviation of the training data) was used to ensure that all input variables get weighted appropriately

In the context of model optimization and system identification, several frameworks aim to extract meaningful insights from data. One such approach is SINDy [6]. While SINDy is not entirely one-shot (first step of parameter estimation is followed by iteratively sparsifying the coefficients), it comes up with a single best-fit model. AI-DARWIN comes up with multiple options which allow for a broader range of discovery of functional relations between the input variables. This is particularly useful for systems biology models and surrogate models of complex systems. This also can provide a pareto front for further examination of the trade-off between simpler but less accurate models in contrast to complex but more accurate models. Further, AI-DARWIN provides flexibility to incorporate domain-specific constraints, which while limiting in accuracy, reaps rewards in interpretability and eventual robustness over unseen data. The resulting models are expressed in polynomial form, offering a clear interpretation of how each factor influences production outcomes. The generated datasets were used to train the models, with performance assessed using accuracy metrics such as Mean Absolute Error (MAE) and interpretability metrics like R^2.

For each dataset, we conducted a series of

modeling steps to predict the concentrations of lactose (C,LA), adenine (C,AD), and lovastatin (C,LOV) in the final stream of the plant (stream #10). Figure 3 shows all the inputs to the model and outputs. However, in AI-DARWIN, we are instructing AI-DARWIN to review all features, select the ones it deems optimal, and potentially arrive at a reduced set of features for the final model.

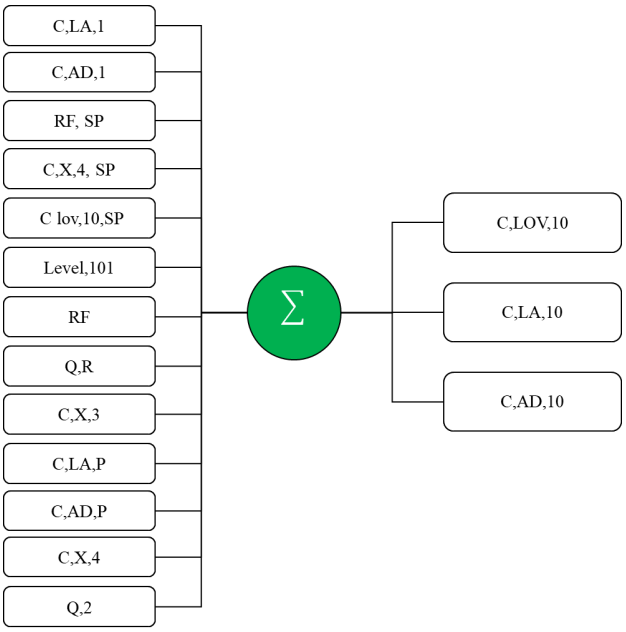

Figure 3. All the inputs and outputs of the AI-Darwin framework (C: Concentration, LA: Lactose, AD: Adenine, X: Biomass, LOV: Lovastatin, SP: Setpoint, RF: Recycle factor, Q: flow rate; the numbers represent the streams number at figure 1).

3. RESULTS AND DISCUSSION

For each dataset several model architectures were used, and the performance is compared. Table 1 and 2 show selected models developed to predict the concentration of two substrates consumed and the API produced at the outlet of the plant. The results demonstrate that the developed modeling framework effectively captures the complex dynamics of the Lovastatin production plant with low training and testing MAE values, demonstrating high predictive accuracy. Using a combination of mechanistic principles and data-driven modeling, the framework successfully predicts three critical outlet concentrations—C,AD, C,LAC, and C,LOV—at the plant's downstream stage.

For Dataset 1, the models showed strong performance and the minimal error difference between training and testing for all the models suggests strong generalization (Table 1 and Figure 4). The polynomial models, expressed in interpretable forms, highlighted significant contributions from key variables such as 'Q,R' and 'C,LA,1' for predicting adenine concentration, and 'Q,R', 'RF', 'C,X,4', and 'C,AD,P' for predicting lactose and lovastatin concentrations. These terms reflect critical process interactions and nonlinearities inherent in this biomanufacturing system. Adenine concentration in stream 10 (C,AD), being the simplest model, achieves remarkable accuracy (MAE_{test}=0.0198), indicating that minimal complexity may suffice for certain targets. In contrast, lactose and lovastatin concentrations (C,LA and C,LOV) include higher-order terms like (C,X,4^4) and (C,AD,P^4), reflecting more intricate relationships.

These terms suggest significant nonlinearity in the system, consistent with the expected process dynamics. The inclusion of interaction terms, such as (C,X,4)(C,AD,P^2), highlights interdependencies between process variables, offering deeper insights into their combined effects on product concentrations. While more complex models (e.g., C,LA and C,LOV) slightly improve accuracy, they may require more computational resources compared to C,AD.

The models developed using Dataset 2 retained this interpretability while incorporating additional interactions and higher-order terms (Table 2 and Figure 5). For example, the C,LA model included interactions like (RF^2),(C,X,4^2), and (Q,R^2), emphasizing the interdependencies between upstream and downstream variables. Similarly, the C,LOV model highlighted the cascading influence of intermediate concentrations, such as C,LA,P on the final product.

The similarity between the MAEs for training and testing sets is due to the relatively large dataset and the low model complexity, which together reduce the risk of overfitting. The large dataset and high sampling

Table 1. Model performance on dataset 1 .

Target	Model	Train MAE	Test MAE
C,AD	$9.72 + 0.25\,(C,LA,1) - 0.56\,(C,LA,1^2)\,(Q,R)$	0.0198	0.0198
C,LA	$75.4 + 6.45\,(C,X,4^4) - 0.60\,(C,X,4)(C,AD,P^2) + 5.17\,(C,AD,P^3) - 0.94\,(Q,R)\,(RF^2)$	0.0785	0.0741
C,LOV	$12.04 + 5.17\,(C,X,4^4) - 0.52\,(C,X,4)(C,AD,P^2) + 6.78(C,AD,P^4) - 0.45\,(Q,R)(RF^2)$	0.0982	0.0971

Table 2. Model performance on dataset 2 .

Target	Model	Train MAE	Test MAE
C,AD	$10.18 - 1.75\,(Q,R) \times (RF^2) + 2.06\,(Q,R^4)$	0.0158	0.0154
C,LA	$77.04 + 3.18\,(C,X,4^4) - 3.21\,(RF^2)(C,X,4^2) + 3.20\,(C,AD,P)\,(Q,R^2) - 2.48\,(C,AD,P^4)$	0.0970	0.0959
C,LOV	$10.43 - 0.90\,(C,X,4^4) - 0.45\,(RF) + 0.41\,(C,LA,P)(C,X,4) - 0.20\,(C,AD,P^4)$	0.0279	0.0277

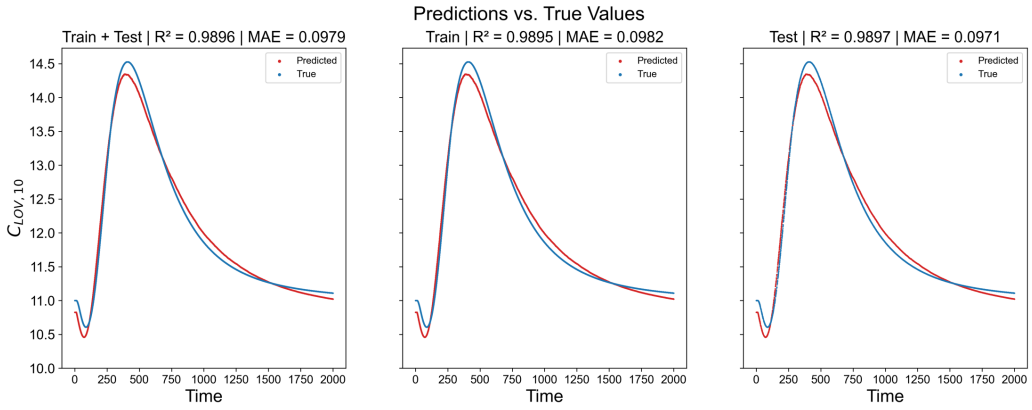

Figure 4 Model performance on the dataset 1

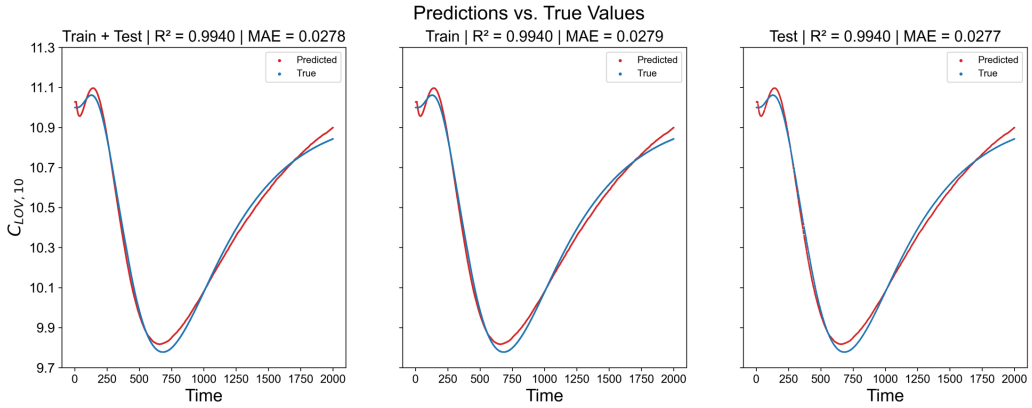

Figure 5. Model performance on the dataset 2

frequency could contribute to the observed similarity between training and testing MAEs. However, this sampling strategy reflects realistic process monitoring where frequent measurements are common. The consistent performance across both subsets suggests that the model captures the underlying process behavior without overfitting. Nevertheless, we agree that further validation under different operating conditions would provide a stronger confirmation of generalizability.

Figure 6 presents the sensitivity analysis performed using MAPE. In Dataset 1, the C,AD,P point's distance from the center of the radar chart, represented by the blue line, indicates the highest sensitivity, meaning minor changes in C,AD,P significantly alter the MAPE. Although C,AD,P remain highly influential in Dataset 2 (magenta line), their impact is slightly reduced, which may be attributed to the overall improved fitting performance of the model. Similarly, C,X,4 exhibited significant sensitivity in both datasets, though to a lesser extent than C,AD,P, with Dataset 1 showing a marginally higher sensitivity than Dataset 2. Conversely, RF, Q,R, and C,LA,P demonstrated considerably lower sensitivity in both datasets. Despite slight variations in magnitude, the overall sensitivity patterns regarding biomass concentration were consistent across both datasets, with Dataset 1 showing a slightly higher sensitivity to changes in C,X,4. However, discrepancies in the sensitivities of RF and C,AD,P may arise from the fact that these represent two scenarios modeled for different data conditions (step changes in lactose and adenine, respectively).

This modeling framework demonstrates its capability to model a plantwide biomanufacturing process, laying the groundwork for developing digital twins. The interpretability of the polynomial models facilitates understanding the underlying system dynamics, while the inclusion of complex interactions and nonlinear terms ensures the accuracy needed to capture the intricacies of the process. For models involving biomass concentration at stream 4 (C,X,4 after the fermentation stage) the kinetic models in the KTB1 model can be replaced to calculate biomass concentration in the CSTR for new microorganisms. By transferring the process dynamics knowledge captured in the polynomial models to new scenarios for different products, modeling of new scenarios can also be performed. This idea will be explored in future work, and validation will be required.

Moreover, in the presented study, step changes in lactose and adenine concentrations—representing typical faults encountered in industrial processes—are introduced to the process. Within a broader digital twin framework that includes a fault diagnosis module, once faults are detected and diagnosed, the resulting changes in Lovastatin concentration at the output can be effectively estimated. This estimation can be done using key

process variables, such as flow rates (directly measurable) or biomass concentration (easily calculable), by the models proposed in this study. The biomass concentration can be approximated through a kinetics model within the CSTR. This approach provides an opportunity for the plant to predict the final product concentration—after several units—using operational parameters (recycling factor and flow rate) along with the biomass concentration measured after the fermenter (via the kinetics model) beside nutrient concentrations after the hydro-cyclone.

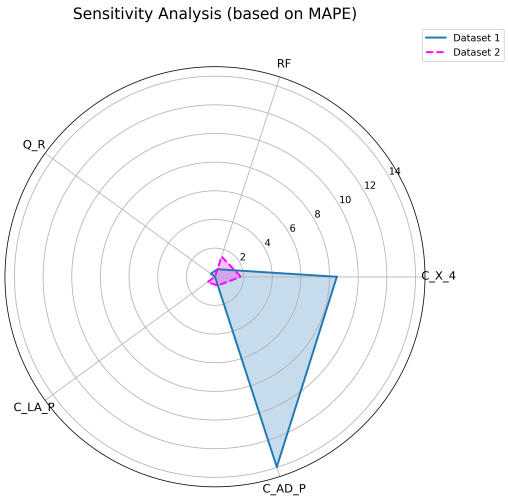

Figure 6. Sensitivity analysis of the models developed for lovastatin production using both datasets.

4. CONCLUSION

The polynomial models developed in this study accurately capture the complex dynamics of a plantwide bio-manufacturing process, making the framework a valuable tool for advancing digital twins. By incorporating higher-order and interaction terms, the models effectively represent nonlinear relationships and interactions between process variables, enabling predictions under varying operational conditions.

The low MAE values for both training and testing datasets indicate that the models generalize well, balancing accuracy and robustness. Their explicit functional forms provide interpretability, allowing for insight into variable sensitivities and interactions, which is essential for developing digital twins.

Future work will focus on generating more complex datasets by varying multiple variables simultaneously, providing a richer dataset for model development. Additionally, testing more advanced models to ensure the selection of the best, yet simplest, representation would further validate this framework. This approach, coupled with a model selection framework, will optimize the trade-off between accuracy, complexity, and interpretability, expanding the applicability of the models across diverse scenarios. By using polynomial models to determine optimal process setpoints, we can adapt the system for different scenarios, such as changing the microorganism to produce a new product, and transfer insights to new contexts. Additionally, generating natural language explanations from the obtained ML plantwide models by incorporating domain-specific knowledge alongside large language models (LLMs) is part of the vision for future works. This approach could lead to the development of a large knowledge model (LKM) [7] tailored for biomanufacturing. This enhances the interoperability of our obtained models and can be used to generate additional mechanistic insights.

REFERENCES

1. Boskabadi, M., Ramin, P., Kager, J., Sin, G., & Mansouri, S. S. (2024). KT-Biologics I (KTB1): A dynamic simulation model for continuous biologics manufacturing. Computers and Chemical Engineering, 108770.
2. Venkatasubramanian, V. (2009). Drowning in data: Informatics and modeling challenges in a data-rich, networked world. AIChE Journal, 55(1), 2-8.
3. Shahhoseyni, S., Greco, L., Sivaram, A., & Mansouri, S. S. (2024). A reduced-order hybrid model for photobioreactor performance and biomass prediction. Algal Research, 84, 103750.
4. Chakraborty, A., Serneels, S., Claussen, H., & Venkatasubramanian, V. (2022). Hybrid AI models in chemical engineering – A purpose-driven perspective. Computer Aided Chemical Engineering, 51, 1507-1512. Elsevier.
5. Chakraborty, A., Sivaram, A., & Venkatasubramanian, V. (2021). AI-DARWIN: A first principles-based model discovery engine using machine learning. Computers & Chemical Engineering, 154, 107470.
6. Wentz, J., & Doostan, A. (2023). Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data. Computer Methods in Applied Mechanics and Engineering, 413, 116096.
7. Venkatasubramanian, V., & Chakraborty, A. (2025). Quo Vadis ChatGPT? From large language models to large knowledge models. Computers & Chemical Engineering, 192, 108895.