Confronting knowledge-based and machine learning models in describing batch fermentation.

Núria Campo-Manzanares[*,**] **Artai Rodríguez-Moimenta**[*,**] **Romain Minebois**[***] **Amparo Querol**[***] **Eva Balsa-Canto**[*,†]

[*] *Bioprocess and Biosystems Engineering, IIM-CSIC, Vigo, Spain.*
[**] *Applied Mathematics Dept. II, University of Vigo, Spain.*
[***] *YeastOmics, IATA-CSIC, Valencia, Spain*

[†] Corresponding author: ebalsa@iim.csic.es

1. INTRODUCTION

Batch fermentation processes are widely used in industry to produce antibiotics, enzymes, biofuels, and fermented foods and beverages such as wine, yogurt, bread, and beer. The underlying concept is to introduce specific microorganisms into a medium with the nutrients necessary for cells to grow. During growth, microorganisms transform nutrients into biomass, releasing the desired products.

Optimizing fermentation processes, including species and conditions, is essential for improving yield and productivity. Knowledge-based models facilitate decision making while minimizing experiments (Lopatkin and Collins, 2020; Wang et al., 2023). Although they offer many advantages, the formulation of such models requires data, time, and insight.

In recent years, machine learning (ML) algorithms have gained significant attention due to their ability to analyze vast amounts of data and uncover patterns that traditional methods might miss. ML has the potential to optimize workflows, reduce costs, and improve product quality. In the context of fermentation, ML has been used to predict production when historical data are available (Shah et al., 2022) or as a surrogate model for scaling up (del Rio-Chanona et al., 2019). However, its effectiveness relies on the availability and quality of the data.

Fermentation knowledge-based models are often formulated using time-series data for biomass, substrate, and product dynamics, with fewer than ten sampling points. Experiments may vary temperature or pH to explain their impact. The laws of mass and energy conservation compensate for the limited data. Can ML be applied under these conditions?

This work addresses this question by considering a case study related to yeast fermentation. We first built a knowledge-based model to describe the process under temperature-varying conditions and then used the same data to formulate an ML model of the process.

⋆ This work was funded by MCIN/AEI/10.13039/501100011033 and EU NextGenerationEU/PRTR grant PLEC2021-007827, and Xunta de Galicia (IN607B 2023/04).

Our results show that: i) building a knowledge-based model is an iterative, time-consuming process; ii) ML model design is easier, needing no specific process knowledge, but requires testing multiple architectures; iii) ML models simulation is faster; but iv) ML is not competitive for the same amount of data, offering worse predictive capabilities.

2. RESULTS

2.1 Kinetic model

We have generalized the model by Moimenta et al. (2023) to account for the effect of temperature. The model consists of a set of ordinary differential equations (ODEs) describing biomass growth phases, uptake of sugars and yeast assimilable nitrogen, and relevant products.

The model was built using a multi-experiment identification approach. Six experiments, performed at three different constant temperatures with two different levels of sugars, were used for model formulation and calibration; and two additional experiments, performed at a time-dependent temperature profile were used for validation. We considered five different sets of fermentations led by five industrial yeast species to test the generalizability of the model. Model identification was implemented in the AMIGO2 toolbox (Balsa-Canto et al., 2016).

The model successfully explained the data, with a normalized mean square error in the prediction of less than 10% for all species.

2.2 Machine learning models

We used regression models based on artificial neural networks (ANN), specifically the multilayer perceptron (MLP). We first consider the case of a particular yeast strain (*S. cerevisiae* GALA) and try to predict the production of ethanol, glycerol, acetate, and succinate. We compared three different scenarios:

(1) `Model ML-KIn`: using the same inputs as the kinetic model, including time, initial conditions of temperature and sugar, and yeast assimilable nitrogen (YAN). The model architecture includes 4 input variables, a hidden layer with 2 neurons, and 4 output variables.

DOI: 10.34726/9013

(2) `Model ML-12In`: using twelve inputs, including time, initial conditions, YAN, and the following amino acids: cysteine, glycine, histidine, methionine, aspartate, phenylalanine, isoleucine, and leucine. Note that selected amino acids present distinct dynamic profiles. The model architecture includes 12 input variables, a hidden layer with 3 neurons, and 4 output variables.

(3) `Model ML-22In`: using 22 inputs including time, initial conditions, and all amino acids measured experimentally, except proline, ornithine. The model architecture includes 22 input variables, a hidden layer with 5 neurons, and 4 output variables.

The selected network structures demonstrated minimal overfitting after testing various combinations with the dataset. We also explored combining data from five industrial yeast *S. cerevisiae* strains, which, despite similar ethanol yields, produced varying amounts of glycerol and succinate, enriching the data. The resulting models, `ML-KI-AllSp`, `ML-12I-AllSp`, and `ML-22I-AllSp`, shared the architecture with those obtained with a single species dataset.

For the training of ML models, the input and output data were normalized; missing data was imputed, and outliers were removed from the dataset. The six models were trained using the mean square error (MSE) metric as the loss function, a learning rate of 0.1, a sigmoid activation function and the Stochastic Gradient Descent optimizer. The modeling workflow was implemented in Python using the Keras toolbox (Chollet et al., 2015).

2.3 Comparative analysis

Our results show that when used under the same conditions, ML offers poor performance. Only when data for multiple species were combined, the ML became more accurate (see Figure 1). `Model ML-KI-AllSp` shows an overall normalized mean square error of 18%, while `Model ML-12I-AllSp` shows a 13%, attributed to the increase in input data from 4 to 12 inputs. The addition of data in `Model ML-22I-AllSp` did not result in further improvements. Even with five times more data, the top ML model underperformed compared to the kinetic model.

3. CONCLUSION

The widespread enthusiasm for machine learning (ML) has led to its use in numerous fields. Although ML provides powerful tools for modeling, balancing this excitement with a clear understanding of its limitations and the contexts in which it can truly add value is essential.

In this work, we have confronted knowledge-based kinetic models with ML models in the prediction of yeast batch fermentation. Our results showed that the kinetic model outperformed the ML models, despite the latter being trained on a larger dataset. This is attributed to the fact that ML models rely solely on experimental data and lack prior knowledge, making them susceptible to errors and bias. This highlights the need for a hybrid approach that combines ML and knowledge-based models to exploit their individual advantages and compensate for their individual limitations (Procopio et al., 2023).

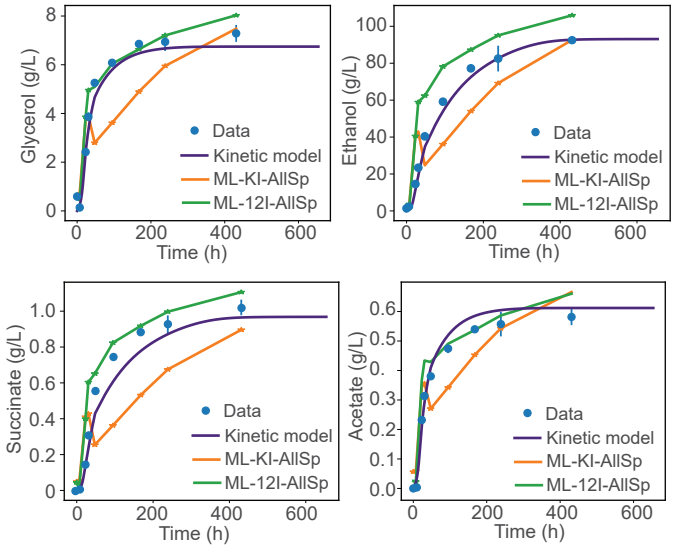

Fig. 1. Kinetic versus ML models for selected examples.

ACKNOWLEDGEMENTS

Funded by MCIN/AEI/10.13039/501100011033 and the EU NextGenerationEU/PRTR grant PLEC2021-007827 and Xunta de Galicia (IN607B 2023/04).

REFERENCES

Balsa-Canto, E., Henriques, D., Gabor, A., and Banga, J. (2016). AMIGO2, a toolbox for dynamic modeling, optimization and control in systems biology. *Bioinformatics*, 32(21), 3357–3359.

Chollet, F. et al. (2015). Keras. `https://keras.io`.

del Rio-Chanona, E.A., Wagner, J.L., Ali, H., Fiorelli, F., Zhang, D., and Hellgardt, K. (2019). Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design. *AIChE J.*, 65(3), 915–923.

Lopatkin, A.J. and Collins, J.J. (2020). Predictive biology: modelling, understanding and harnessing microbial complexity. *Nat. Rev. Microbiol.*, 18(9), 507–520.

Moimenta, A.R., Henriques, D., Minebois, R., Querol, A., and Balsa-Canto, E. (2023). Modelling the physiological status of yeast during wine fermentation enables the prediction of secondary metabolism. *Microb. Biotechnol.*, 16(4), 847 – 861.

Procopio, A., Cesarelli, G., Donisi, L., Merola, A., Amato, F., and Cosentino, C. (2023). Combined mechanistic modeling and machine-learning approaches in systems biology–a systematic literature review. *Comp M Prog Biomed*, 240, 107681.

Shah, P., Sheriff, M.Z., Bangi, M.S.F., Kravaris, C., Kwon, J.S.I., Botre, C., and Hirota, J. (2022). Deep neural network-based hybrid modeling and experimental validation for an industry-scale fermentation process: Identification of time-varying dependencies among parameters. *Chem Eng J*, 441, 135643.

Wang, X., Mohsin, A., Sun, Y., Li, C., Zhuang, Y., and Wang, G. (2023). From Spatial-Temporal Multiscale Modeling to Application: Bridging the Valley of Death in Industrial Biotechnology. *Bioeng.*, 10(6), 744.