|  |  |
| --- | --- |
| cetlogo ***CHEMICAL ENGINEERING TRANSACTIONS***  ***VOL. xxx, 2025*** | A publication of  aidiclogo_grande |
| The Italian Association  of Chemical Engineering  Online at www.cetjournal.it |
| Guest Editors: David Bogle, Flavio Manenti, Piero Salatino  Copyright © 2025, AIDIC Servizi S.r.l. **ISBN** 979-12-81206-21-2; **ISSN** 2283-9216 | |

Dynamic Hybrid Model for biopesticides production using industrial substrate

Sergio Figueroa-Cardona, Carlos E. Robles-Rodriguez, Luc Fillaudeau, Rim El-Jeni, Cesar A. Aceves-Lara\*.

aTBI, Université de Toulouse, CNRS, INRAE, INSA, 135 avenue de Rangueil, Toulouse, France.

aceves@insa-toulouse.fr

*Bacillus thuringiensis* (*Bt*) is the most used microorganism in biopesticides around the world. *Bt* insecticidal activity is primarily due to the release of a protein called δ-endotoxin during sporulation. Industrial production of this biopesticide is challenging to control. A modelling framework could be used to enhance industrial production of these proteins. Several mechanistic models have been developed based on laboratory experiments with synthetic media, and previous works introduced a Dynamic Hybrid Model (DHM) to simulate *Bt* fermentation using Support Vector Machine (SVM) regression in MATLAB®. The goals of this study are to extend the DHM to *Bt* fermentations using industrial substrate and to test other data-driven algorithms. In this context, Decision Tree (DT) regression and SVM are compared. Data from two experiments using one strain of *Bt* was used to train and validate the model. The best DHMs corresponded to SVM with normalized errors (NRMSE) of 0.005 and 0.055 for proteins and spores, respectively. Future research will focus on incorporating additional experimental data to enhance model accuracy. The DHM will be further integrated into a scaled-up process to optimize industrial biopesticide production.

Keywords: hybrid modelling, dynamic modelling, bioprocess, Support Vector Machine, Decision Tree.

* 1. Introduction

Biopesticides are a strong alternative to reduce the cost and environmental impact of synthetic pesticides (Tadesse Mawcha *et al.*, 2024). Most commercial biopesticides of microbial origin are based on *Bacillus thuringiensis (Bt)*, a Gram-positive sporulating bacteria, that produces δ-endotoxins that are toxic against a wide spectrum of insects, such as lepidopterans, coleopterans and dipterans (Rowe & Margaritis, 2004). These δ-endotoxins are produced in the form of crystalline inclusions, also called Cry-proteins. *B. thuringiensis* must experience certain modifications of its physiology to produce these proteins which makes its culture and monitoring difficult. One alternative to optimize the production of these δ-endotoxins is to use a model-based approach in which multiple simulations of mathematical models are assessed. There are only a few models available that can correctly estimate the dynamics of *B. thuringiensis growth* and sporulation phases (Navarro-Mtz & Pérez-Guevara, 2014). However, this model was built for ad-hoc experiments and thus its extrapolation to other data is not suitable as it depends on a large number of parameters for specific conditions.

Dynamic Hybrid Modelling (DHM) integrates mechanistic and data-driven models, proposing cost-effective solutions to model biochemical processes that are complex and whose underlying mechanisms are not well known (Shah *et al*., 2025). Even though the construction of a dynamic hybrid model still requires a large collection of data through traditional experiments for recently developed processes, it has proven to be more efficient using data than purely data-driven methods (Kay *et al*., 2023).

Hybrid modelling has been applied in fermentation-related bioprocesses to compare the performance of kinetic models and their hybrid counterpart when describing the production of a compound of interest (Vega-Ramon *et al*., 2021). In the case of *B. thuringiensis*, a dynamic hybrid model was proposed for cultures carried out under Semi-synthetic Media (SSM) conditions (Figueroa-Cardona *et al*., 2024). This approach consisted in a mass balance model for the most important variables of the process (i.e., biomass, substrate, protein, spores) with the kinetics rates of protein and spores defined by a data-driven model. Support Vector Machine (SVM) was the data-driven method chosen to integrate the hybrid model, based on its versatility when working with small data sets (Vapnik *et al*., 1996). Previous work showed that the implementation of this approach could lead to an improvement in protein and spore prediction. In this work, the dynamic model is first extended and a test on different data-driven algorithms is performed for the prediction of protein and spore production. Additionally, SVM is compared against Decision Tree (DT) regression, as it presents a clear and large structure that branches out on a set of rules based on the predictors (Kharait *et al*., 2007).

The main goal of this paper is to illustrate and apply an improved dynamic hybrid model framework to predict protein and spore production in *B. thuringiensis* fermentation using an industrial substrate. The remainder of this paper is as follows: the dynamic hybrid model is detailed in section 2, results are presented in section 3, and conclusions are drawn in section 4.

* 1. Methods

This section describes the recovery of experimental data; the literature models and the dynamic hybrid model proposed in this work.

* + 1. Organism and culture conditions

Experiments were carried out on a 2 L batch reactor at 30°C with the strain *B. thuringiensis subsp kurstaki* *BLB*1 and wheat bran (WB) as used as substrate. Samples were collected every 2 hours for the first 24 hours and then every 3h for the rest of the culture. Aeration was set to 0.36 L.min-1.

* + 1. Experimental data and data treatment

Experimental data for protein and spore were obtained from two batches, each one with 17 instances (12 samples up to 24h and 5 up to 48h). These data are filtered to improve the quality of the predictors and used to train the data-driven model (machine learning model). A moving median filter was used to understand the data trend and identify outliers. As the experiments have been conducted under batch conditions, it is assumed that the concentration of the products must increase or remain stable, hence it cannot decrease over time. Once the data have gone through the moving median filter, the data have been verified to confirm that every concentration point in a certain time *t* was greater than all the precedent points.

A polynomial function was then fitted to the filtered data through Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) (Fritsch & Carlson, 1980). Additional instances were generated using this function as the target variable of the data-driven models. For model training, 501 instances were generated per variable. However, for spores, only 424 were used due to the deletion of null values (spore concentration is zero at the beginning of the culture).

* + 1. Dynamic Model for industrial substrate

An extended dynamic model (DM) was proposed based on the model of Monroy *et al.* in 2021. The mass balances for *B. thuringiensis* in a batch reactor for industrial substrate as described in Eq. (1 - 6).

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  | (3) |
|  | (4) |
|  | (5) |
|  | (6) |

Where X refers to biomass, W to wheat bran, G to glucose, M to an intermediate compound, P for protein, all variables expressed in g.L-1, and Spo for the spore concentration, in units of CFUx10-5.L-1.

The biomass growth rate is represented by µ and kd is the biomass decay constant, both in units of h-1. The Contois expression was used to calculate the biomass growth rate, as given in Eq. (7). The constant µmax is the maximal growth rate (h-1) and KC is the Contois specific constant (gG.gX-1).

|  |  |
| --- | --- |
|  | (7) |

The model proposed in Monroy *et al.* (2021) was extended to consider wheat bran (W) as main substrate. The hydrolysis of W is a complex reaction that considers an intermediate compound (M) that hydrolyzes into glucose (Bednarska, 2015) as:

|  |  |
| --- | --- |
|  | (8) |

To consider the influence of these compounds on glucose production, a mass balance for species M is added to dynamic model in eq.(3). In this equation, parameter YWM describes the yield of production of M from W hydrolysis. The kinetics to describe the specific reaction rates are described by qW (gW.gX-1.h-1) as in eq.(9) for W to M, while the specific reaction rate from M to G is represented by qM (gM.gX-1.h-1) as in eq.(10),

|  |  |
| --- | --- |
|  | (9) |
|  | (10) |

These equations, based on Contois kinetic laws, rely on parameters k1 (gW.gX1.h-1) and k2 (gW.gX-1) for W consumption, and parameters k3 (gM.gX-1.h-1) and k4 (gM.gX-1) for M consumption.

Parameters YXG and YMG denote the yield coefficient between biomass and glucose (gX.gG-1) and the mass yield of M consumed per glucose produced (gM.gG-1). Parameter mG (gG.gX-1.h-1) considers the glucose consumption for maintenance.

The parameters kpro and kspo are specific kinetic constants for protein and spore in gPro.gX-1.h-1 and CFUx10-5. gX-1.h-1, respectively. Spore production is triggered by a ratio of carbon over nitrogen (C/N) that states when sporulation begins. This is because spores are produced under stress conditions when the substrate is limiting (Amicarelli *et al*., 2010). An approximation of this C/N ratio was obtained for the fermentation conditions as a glucose to biomass ratio (G/X) calculated previously in *Bt* experiments that used semi-synthetic media. The constant φspo is defined as the threshold value of G/X that induces the sporulation.

* + 1. Dynamic Hybrid Model

The proposed Dynamic Hybrid Model (DHM) aims to estimate the kinetic rates of protein and spore by using data-driven models. This means that these rates will be re-estimated based on the dynamic model and other predictors to improve the prediction of protein and spore. Hence, Eq.(5) and (6) are replaced by,

|  |  |
| --- | --- |
|  | (11) |
|  | (12) |

The output variables, r\*p and r\*spo, were obtained using the numerical derivative of the filtered data for protein and spore, respectively. Two data-driven algorithms are proposed for comparison: Decision Tree (DT) and Support Vector Machine (SVM). Decision Trees (DT) are structured in a hierarchical scheme that classifies data, sometimes partitioned, into larger uniform subsets based on a series of logical decisions (Kharait *et al*., 2007). This tree-based method accommodates both categorical and numerical data, delivering clear and interpretable results (Kavoni *et al*., 2025). All the models were trained using the regression learner tool from MATLAB® 2020a. The model types were chosen based on software availability. Three model types were tested for DT: Fine, Medium, and Coarse, based on the minimum leaf size of the type of model.

On the other hand, SVM is a non-parametric, non-linear technique that has gained significant attention in recent years for its stability, robustness, and versatility, particularly in high-dimensional regression and classification tasks (Hamedi *et al*., 2021). In this work, SVM is used to predict accurately the desired output through Eq (13).

|  |  |
| --- | --- |
|  | (13) |

where φ(x) is the nonlinear mapping of the input *x* into a high dimensional feature space. Different kernels were tested: Linear, Quadratic, Cubic, and Gaussian. The hyperparameters of both data driven methods were optimized. A Bayesian optimization limited to 150 iterations was performed for each case. The optimized model for DT corresponded to a Fine model, while a Gaussian kernel function was used for SVM.

The data driven models for r\*p and r\*spo can be represented as a function of the form:

|  |  |
| --- | --- |
|  | (14) |

Both models were defined based on four predictors: biomass growth rate (µ) from Eq. (7), and the consumption or production rates from Eq. (4) for glucose (dG/dt), Eq. (5) for proteins (rp) and Eq. (6) for spores (rspo). The results of the implementation of these models coupled with the dynamic model are presented in the results section.

* 1. Results

The data-driven models, that belong to the Hybrid model, were trained using data from the batch that reported a wider range of concentration for each variable. Batch B2 was selected to train the protein rate, and batch B1 in the case of spore. In order to compare the performance of the Dynamic model (DM) against the Dynamic Hybrid model (DHM), the complete models were assessed regarding protein and spore concentrations. The predictions were assessed through the mean absolute error (MAE), root mean squared errors (RMSE) and the normalized root mean squared errors (NRMSE) of the models compared to the filtered data in Table 1.

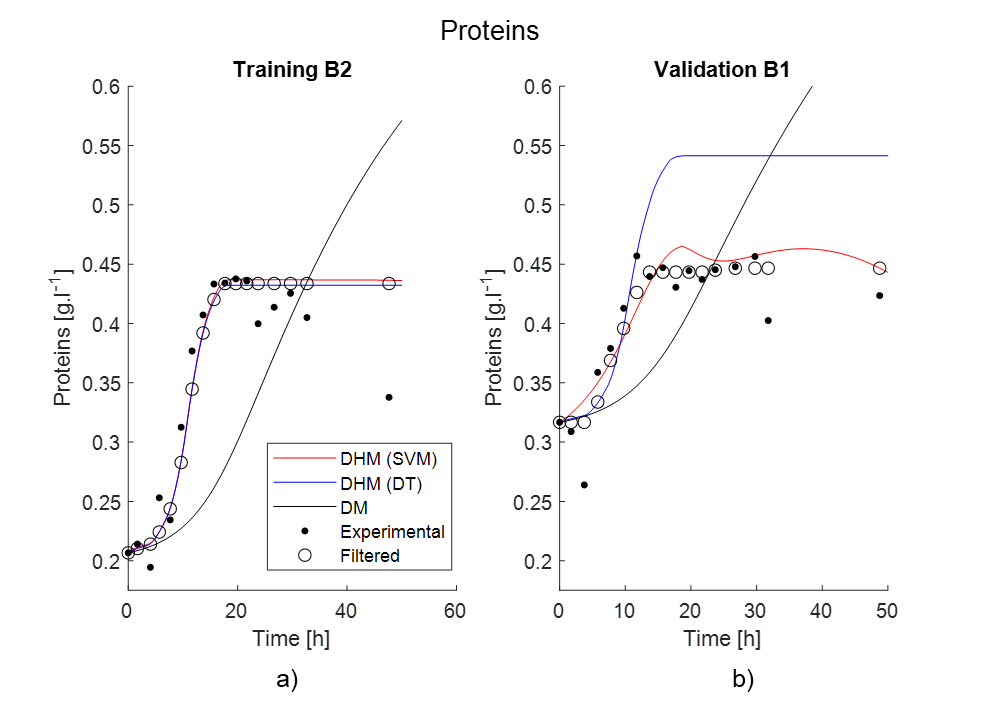
Regarding protein concentration, DHMs with the lowest error for the three metrics for the training were Support Vector Machine (SVM) with Gaussian kernel function, and the Fine Decision Tree (DT) with optimized leaf size (Fine (Opt)). Although the NRMSE is the same for SVM Gaussian and Gaussian (Opt), the first method was preferred based on the training time of the data-driven model (0.8 s and 699 s, respectively). The DHM with the lowest error for validation test was SVM with Linear kernel function. However, all DT methods show equivalent metrics results, where Medium DT was slightly better. Figure 1 shows the comparison between the protein concentration estimated using the DM and the DHMs with the lowest error for each batch.

Table 1: Error comparison of training and validation tests for protein models using filtered data.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test type |  | Training (B2) | | | Validation (B1) | | |
| Method | Type | MAE | RMSE | NRMSE | MAE | RMSE | NRMSE |
| DM | - | 7.60 x10-2 | 9.54 x10-2 | 0.202 | 5.51 x10-2 | 7.69 x10-2 | 0.213 |
| DHM (DT) | Fine | 2.17 x10-3 | 2.49 x10-3 | 0.005 | 6.51 x10-2 | 7.85 x10-2 | 0.218 |
| Medium | 2.53 x10-3 | 2.97 x10-3 | 0.006 | **6.07 x10-2** | **7.32 x10-2** | **0.203** |
| Coarse | 5.64 x10-3 | 8.33 x10-3 | 0.018 | 6.69 x10-2 | 7.93 x10-2 | 0.220 |
| Fine (Opt) | **1.00 x10-3** | **1.16 x10-3** | **0.002** | 6.17 x10-2 | 7.43 x10-2 | 0.206 |
| DHM (SVM) | Linear | 5.84 x10-2 | 6.78 x10-2 | 0.144 | **1.10 x10-2** | **1.24 x10-2** | **0.034** |
| Quadratic | 4.48 x10-3 | 5.65 x10-3 | 0.012 | 5.91 x10-2 | 7.12 x10-2 | 0.198 |
| Cubic | 6.86 x10-2 | 9.82 x10-2 | 0.208 | 1.29 x10-1 | 1.67 x10-1 | 0.463 |
| Gaussian | **1.92 x10-3** | **2.29 x10-3** | **0.005** | 5.89 x10-2 | 7.09 x10-2 | 0.197 |
| Gaussian (Opt) | 1.95 x10-3 | 2.23 x10-3 | 0.005 | 6.56 x10-2 | 7.85 x10-2 | 0.218 |

Table 2: Error comparison of training and validation tests for spore models using filtered data.

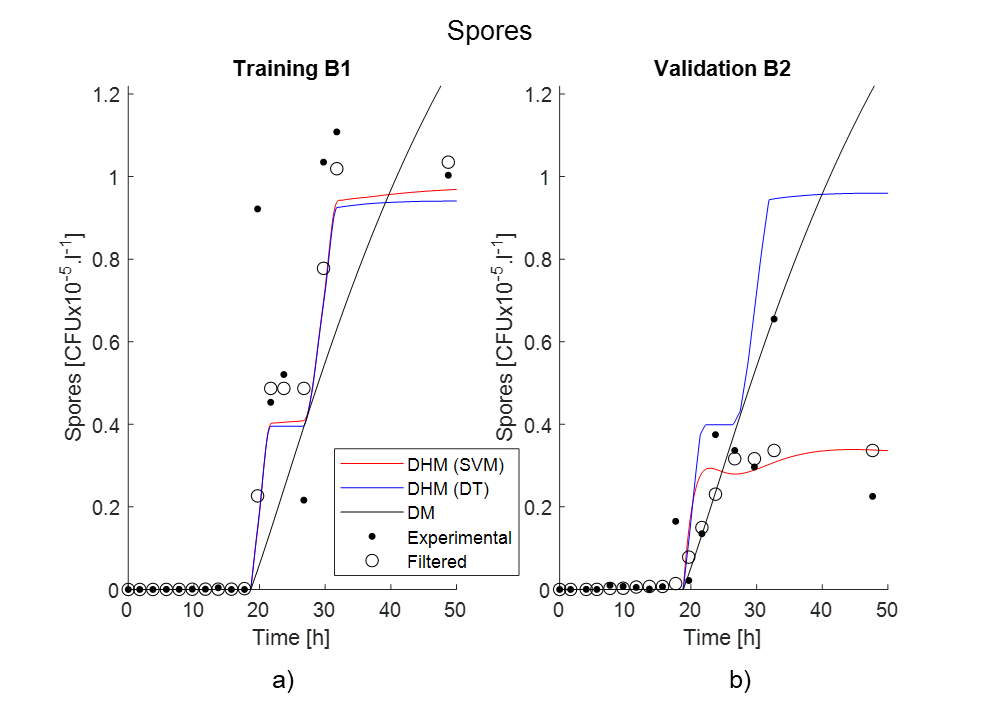
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Test type |  | Training (B1) | | | Validation (B2) | | |
| Method | Type | MAE | RMSE | NRMSE | MAE | RMSE | NRMSE |
| DM | - | 1.22 x10-1 | 1.84 x10-1 | 0.181 | 1.05 x10-1 | 2.48 x10-1 | 0.427 |
| DHM (DT) | Fine | 4.72 x10-2 | 6.66 x10-2 | 0.066 | 1.49 x10-1 | 2.62 x10-1 | 0.452 |
| Medium | 4.60 x10-2 | 6.48 x10-2 | 0.064 | **1.44 x10-1** | **2.56 x10-1** | **0.441** |
| Coarse | 5.14 x10-2 | 7.37 x10-2 | 0.072 | 1.49 x10-1 | 2.71 x10-1 | 0.468 |
| Fine (Opt) | **4.51 x10-2** | **6.33 x10-2** | **0.062** | 1.47 x10-1 | 2.60 x10-1 | 0.448 |
| DHM (SVM) | Linear | 1.87 x10-1 | 3.18 x10-1 | 0.312 | **2.62 x10-2** | **4.37 x10-2** | **0.075** |
| Quadratic | 1.70 x10-1 | 2.64 x10-1 | 0.259 | 6.28 x10-2 | 1.14 x10-1 | 0.197 |
| Cubic | 5.70 x10-1 | 8.51 x10-1 | 0.837 | 4.23 x10-1 | 7.28 x10-1 | 1.254 |
| Gaussian | 6.56 x10-2 | 9.37 x10-2 | 0.092 | 1.22 x10-1 | 2.22 x10-1 | 0.383 |
| Gaussian (Opt) | **4.00 x10-2** | **5.62 x10-2** | **0.055** | 1.42 x10-1 | 2.53 x10-1 | 0.437 |



*Figure 1:* *Model comparison for protein concentration. a) best training (Fine Opt DT / Gaussian SVM); b) best validation (Medium DT / Linear SVM).*

The error metrics for spore concentration are presented in Table 2. The best DHMs were SVM using Gaussian kernel with optimized hyperparameters (Gaussian (Opt)) and Fine (Opt) DT. The best DHMs estimation using the training data are satisfactory compared to the DM, mostly considering the large changes in spore concentration. On the contrary, simulation in validation batches shows a gap when compared to experimental results, as depicted in Figure 2. The methods depicted for validation test were the same as for protein concentration (Medium DT and SVM Linear kernel), even though there is no clear difference between the error metrics between DT methods.

The best SVM DHMs error in both validation batches was considerably minor compared to the initial prediction made by the DM. The difference between the best methods in DHM for training and validation tests is due to the discrepancy of experimental data between both batches. The SVM using Linear kernel is more accurate compared to the DM when describing the variables concentration in the validation batches, achieving a NRMSE reduction of 18% and 35% for protein and spore, respectively.



*Figure 2:* *Model comparison for spore concentration. a) best training (Fine Opt DT / Gaussian Opt SVM); b) best validation (Medium DT / Linear SVM).*

* 1. Conclusions

As the *B. thuringiensis* dynamics are complex and unknown, mechanistic models cannot correctly estimate the protein and spore production from its fermentation. The proposed dynamic hybrid model improved the prediction of protein and spore even when the kinetics of the dynamic model were neither well defined nor identified. Results show that SVM proved to be more accurate than DT. However, a larger improvement could be obtained if more data is used for training and validation. The DHM will be applied to optimize industrial production of biopesticides in a process chain framework with different *B. thuringiensis* strains.

Acknowledgments

This work was supported by the “Alternative Biopesticides For Safe Integrated Pest and Water Management Around Mediterranean” (SAFWA) project, which has received funding from the European Union’s Horizon 2020 research and innovation program under the grant agreement No2022/Section 2. This work was partially supported by the French National Research Agency (ANR) under the "Investissements d'Avenir" program, reference ANR-18-EURE-0021.

References

Amicarelli, A., Sciascio, F., Quintero, O., & Ortiz, O. A. (2010). *On-line Biomass Estimation in a Batch Bio-technological Process: Bacillus Thuringiensis δ - Endotoxins Production* (pp. 79–112). https://doi.org/10.5772/9769

Bednarska, K. A. (2015). *Kinetic modelling of enzymatic starch hydrolysis*. Wageningen University.

Figueroa-Cardona, S., Robles-Rodriguez, C. E., El-Jeni, R., Fillaudeau, L., & Aceves Lara, C. A. (2024). Dynamic hybrid modelling for biopesticides production using Bacillus Thuringiensis strains. *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-3-031-82073-1\_24

Fritsch, F. N., & Carlson, R. E. (1980). Monotone Piecewise Cubic Interpolation. *SIAM Journal on Numerical Analysis*, *17*(2), 238–246. https://doi.org/10.1137/0717021

Hamedi, H., Mohammadzadeh, O., Rasouli, S., & Zendehboudi, S. (2021). A critical review of biomass kinetics and membrane filtration models for membrane bioreactor systems. *Journal of Environmental Chemical Engineering*, *9*(6), 106406. https://doi.org/10.1016/j.jece.2021.106406

Kavoni, H., Savizi, I. S. P., Lewis, N. E., & Shojaosadati, S. A. (2025). Recent advances in culture medium design for enhanced production of monoclonal antibodies in CHO cells: A comparative study of machine learning and systems biology approaches. *Biotechnology Advances*, *78*, 108480. https://doi.org/10.1016/j.biotechadv.2024.108480

Kay, S., Kay, H., Rogers, A. W., & Zhang, D. (2023). Integrating hybrid modelling and transfer learning for new bioprocess predictive modelling. In A. C. Kokossis, M. C. Georgiadis, & E. Pistikopoulos (Eds.), *Computer Aided Chemical Engineering* (Vol. 52, pp. 2595–2600). Elsevier. https://doi.org/10.1016/B978-0-443-15274-0.50412-1

Kharait, S., Hautaniemi, S., Wu, S., Iwabu, A., Lauffenburger, D. A., & Wells, A. (2007). Decision tree modeling predicts effects of inhibiting contractility signaling on cell motility. *BMC Systems Biology*, *1*(1), 9. https://doi.org/10.1186/1752-0509-1-9

Monroy, T. S., Abdelmalek, N., Rouis, S., Kallassy, M., Saad, J., Abboud, J., Cescut, J., Bensaid, N., Fillaudeau, L., & Aceves-Lara, C. A. (2021). Dynamic Model for Biomass and Proteins Production by Three Bacillus Thuringiensis ssp Kurstaki Strains. *Processes*, *9*(12), 2147. https://doi.org/10.3390/pr9122147

Navarro-Mtz, A. K., & Pérez-Guevara, F. (2014). Construction of a biodynamic model for Cry protein production studies. *AMB Express*, *4*(1), 79. https://doi.org/10.1186/s13568-014-0079-y

Rowe, G. E., & Margaritis, A. (2004). Bioprocess design and economic analysis for the commercial production of environmentally friendly bioinsecticides from Bacillus thuringiensis HD-1 kurstaki. *Biotechnology and Bioengineering*, *86*(4), 377–388. https://doi.org/10.1002/bit.20146

Shah, P., Pahari, S., Bhavsar, R., & Kwon, J. S.-I. (2025). Hybrid modeling of first-principles and machine learning: A step-by-step tutorial review for practical implementation. *Computers & Chemical Engineering*, *194*, 108926. https://doi.org/10.1016/j.compchemeng.2024.108926

Tadesse Mawcha, K., Malinga, L., Muir, D., Ge, J., & Ndolo, D. (2024). Recent Advances in Biopesticide Research and Development with a Focus on Microbials. *F1000Research*, *13*, 1071. https://doi.org/10.12688/f1000research.154392.3

Vapnik, V., Golowich, S., & Smola, A. (1996). Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. *Advances in Neural Information Processing Systems*, *9*. https://proceedings.neurips.cc/paper\_files/paper/1996/hash/4f284803bd0966cc24fa8683a34afc6e-Abstract.html

Vega-Ramon, F., Zhu, X., Savage, T. R., Petsagkourakis, P., Jing, K., & Zhang, D. (2021). Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty. *Biotechnology and Bioengineering*, *118*(12), 4854–4866. https://doi.org/10.1002/bit.27950