From Ethanol to Higher Alcohols: Unlocking the Economics of Guerbet Coupling Through Machine Learning

Guilherme Vieira Espinosa[a], Letícia Costa da Silva Mesquita[a], Leonardo Dantas de Souza Netto[a], Amanda Lemette Teixeira Brandão [a,*]

[a] *Pontifical Catholic University of Rio de Janeiro, Department of Chemical and Materials Engineering, Rio de Janeiro-RJ, Brazil*
[*] *amanda.lemette@puc-rio.br*

ABSTRACT

The conversion of ethanol into higher alcohols via Guerbet coupling is a promising route for producing advanced biofuels with improved properties. However, assessing the economic feasibility of the processes relies on computationally expensive simulations. The present work proposes data-driven surrogate models to predict the economic performance of an ethanol-to-higher-alcohols plant, reducing computational cost while preserving accuracy. A detailed Aspen Plus simulation was used to generate 10,000 scenarios by varying process and economic parameters and the internal rate of return (IRR) was adopted as the performance metric. Three machine learning models, linear regression, eXtreme Gradient Boosting (XGBoost), and a multilayer perceptron (MLP), were evaluated and compared, along with feature selection techniques to identify the most relevant variables. Nonlinear models outperformed linear regression, with the MLP achieving the best performance (R^2 = 0.998). Feature reduction improved model robustness, and SHapley Additive exPlanations (SHAP) analysis identified ethanol price as the most influential variable, negatively affecting IRR. This framework enables fast and accurate evaluation of process economics while significantly reducing computational effort, offering a practical and scalable alternative to traditional simulation-based techno-economic analysis.

Keywords: Ethanol conversion, Guerbet coupling, Techno-economic evaluation, Machine learning

1 Introduction

The production of liquid biofuels expanded significantly in the late 20th century, with bioethanol and biodiesel emerging as promising alternatives to fossil fuels (Bušić et al., 2018). Ethanol represents 70% of global biofuel production, originated from sugar crops, lignocellulosic biomass, microalgae or genetically engineered feedstocks (Hamden et al., 2025). Although ethanol can be directly used as a fuel or fuel additive, offering a high octane number and reduced greenhouse gas emissions, its hygroscopicity, lower energy density and the need for engine modifications limit its applicability and motivate the investigation of alternative fuels (Zhang et al., 2016). Therefore, upgrading bioethanol to higher alcohols such as n-butanol through catalytic processes constitutes a promising approach to aid in the energy transition. Butanol is used as fuel or fuel additive, offering advantages over ethanol, such as lower water affinity and density closer to gasoline. In addition, it is employed as perfume additive, flavoring agent and solvent (Kozlowski & Davis, 2013).

ABE (acetone, butanol and ethanol) fermentation is the most traditional route for industrial n-butanol production. However, despite renewed interest in recent years, challenges associated with product purification, along with the expansion of fossil-based processes, have reduced its industrial attractiveness (Jin et al., 2011; Zhang et al., 2016). To reduce carbon emissions arising from petrochemical processes, the Guerbet coupling of ethanol to produce n-butanol and other even-carbon-numbered alcohols, such as n-hexanol and n-octanol, represents a promising route for the synthesis of lower-emission biofuels (Gabriëls et al., 2015). As the Guerbet coupling mechanism favors successive coupling reactions, longer chain alcohols such as n-hexanol and n-octanol are also generated and may be marketed as valuable byproducts. Conversely, the high water content in the Guerbet reactor effluent and resulting azeotrope formation can increase capital and operational costs due to the associated separation challenges (Rocha et al., 2019).

Process simulation tools are essential for evaluating complex separation systems, such as the previously described alcohol-water effluents, as well as for assessing the profitability of chemical plants designed around these

systems. However, the slow convergence of simulations involving multiple separation unit operations creates an opportunity for the development of data-driven surrogate models to reduce the computational time required for techno-economic evaluation of chemical plants. The integration of machine learning models with process simulation platforms has recently emerged as a promising strategy for reducing the computational cost associated with large-scale techno-economic evaluation (Ordonez et al., 2026). In this context, the present work aims to evaluate data-driven models, trained on Aspen Plus simulation data, that predict the economic performance of Guerbet coupling plants for the simultaneous production of n-butanol, C_6 alcohols and C_8 alcohols, using azeotropic and extractive distillation strategies for product purification and ethanol recycling. By applying regression-based machine learning (ML) models, the present work evaluates the performance of surrogate models in predicting the internal rate of return (IRR) of the simulated process and demonstrates their ability to reduce computational time without compromising accuracy.

2 Methodology

2.1 Process description and dataset generation

The ethanol conversion plant, simulated in Aspen Plus, consists of the Guerbet coupling reactor operating at atmospheric pressure and 350 °C, modeled as described by Tsuchida et al. (2006). After separating lighter gases, the reactor product is subjected to a separation sequence comprised of an azeotropic distillation tower, with benzene as entrainer, to separate the water and ethanol from the higher alcohols mixture followed by two towers to separate butanol, hexanol and octanol. In order to recover ethanol, an extractive distillation section was also included, recycling an ethanol and water mixture, with the same composition as the feed stream, to the reactor. A simplified flow diagram of the process is shown in Figure 1.

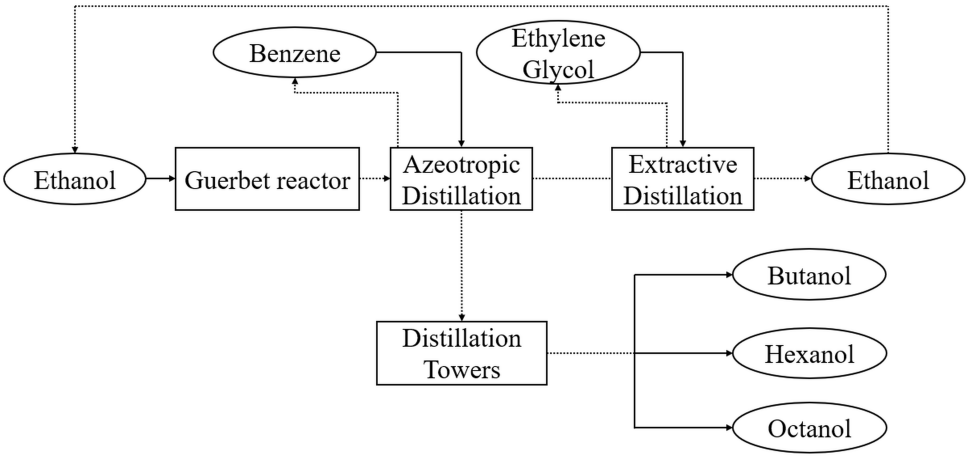

Figure 1: Simplified process flow diagram.

The plant equipment costs were estimated using the Guthrie bare module factors methodology described by Seider et al. (2003) and the operating costs were estimated according to Turton et al. (2009), where costs of raw material, waste treatment, utilities, labor and fixed investment are used as parameters to estimate direct and fixed manufacturing costs. The Guerbet coupling plant profitability was evaluated using internal rate of return (IRR) as performance metric, calculated by solving Equation (1).

$$NPV = \sum_i \frac{CF_i}{(1+IRR)^i} = 0 \tag{1}$$

Where CF_i is the cash flow in year i.

The dataset generation was performed by varying plant capacity, parameters used to estimate total investment (specific equipment's bare module factors, for example) and manufacturing costs, financial parameters such

as federal income taxes and plant lifetime and prices of feedstocks, products and utilities. Overall, 10,000 different combinations of economic parameters were evaluated. A second dataset was derived from the original by replacing all negative IRR values with zero, as variability within the negative region lacks economic significance. The original dataset, however, was employed for both the feature selection and the interpretability analysis stages presented later, to more accurately assess the mathematical influence of the features on IRR prediction.

2.2 *Machine learning workflow*

The IRR prediction problem was framed as a regression task in Python, involving the training and comparative performance evaluation of linear regression, eXtreme Gradient Boosting (XGBoost), and multilayer perceptron (MLP) models across different degrees of model complexity. The generated dataset was split into training and testing sets, comprising 85% and 15% of the data, respectively. The training set was used during the cross-validation stage, employing a K-fold (K = 5) strategy. Within each fold of the cross-validation, min–max normalization was performed on the training subset and applied to the corresponding validation subset. In both the dataset splitting and K-fold cross-validation procedures, a random seed was fixed to guarantee reproducibility of the results across independent runs.

The XGBoost model was trained employing the library's default hyperparameters, while the MLP model was implemented as a fully connected feedforward neural network, with an input layer dimension matching the number of features, followed by two hidden layers with 64 and 32 neurons using ReLU activation function and the output layer consisting of a single neuron with linear activation function. The MLP model was trained using the Adam optimizer with a learning rate of 0.001, employing the mean squared error (MSE), as defined by Equation (2), as the loss function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Where n is the number of samples and y_i and $\hat{y}_i$ are the observed and predicted values for the ith sample.

The predicting performance of the model on the validation and test datasets was primarily evaluated using the mean absolute error (MAE), presented in Equation (3). To complement and provide a more comprehensive assessment, the root mean square error (RMSE) and coefficient of determination (R^2), defined in Equations (4) and (5), were additionally employed.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

Where n is the number of samples, y_i and $\hat{y}_i$ are the observed and predicted values for the ith sample and $\bar{y}$ is the mean of the target variable.

Feature selection was conducted to reduce model complexity by applying two supplementary techniques: least absolute shrinkage and selection operator (LASSO) regression (Tibshirani, 1996), suited for identifying linear relationships between predictors and the target variable, and the XGBoost feature importance (gain) method, developed by Chen and Guestrin (2016), which is capable of capturing nonlinear relationships and interaction effects. Based on the selected subsets of variables, the previously described training and validation procedure was reapplied, now considering a reduced feature space, and the best performance model was evaluated on the test dataset.

To elucidate how each input influences the model's outputs, the SHapley Additive exPlanations (SHAP) approach was utilized, distinguished by its theoretical foundation in Shapley values. The method operates by calculating the marginal contribution of each feature across all possible variable permutations and transforming the model output into a linear sum of local importance values, ensuring an additive and consistent analysis, allowing for the validation of each parameter's influence both in isolation and in combination (Lundberg & Lee, 2017).

3 Results and discussion

3.1 Machine learning model performances

When trained on the complete dataset containing 35 features, the three models exhibited distinct predictive behaviors. In the K-fold cross-validation step, the MLP achieved the best performance (R^2 = 0.996) highlighting its ability to capture complex nonlinear relationships between process and economic variables and the target IRR. XGBoost also showed strong predictive capability, although with slightly lower performance (R^2 = 0.896), while linear regression performed significantly worse (R^2 = 0.172). This result suggests that IRR cannot be adequately represented by a linear combination of the input variables, requiring non-linear models to accurately predict the plant's profitability behavior.

The feature selection procedure identified the following subset of key variables: ethanol, butanol, hexanol, octanol and benzene prices, equity contribution, federal income tax rate and the capital cost of heat exchangers, the latter represented by their bare module (BM) factors (Seider et al., 2003). This dimensionality reduction led to consistent performance improvements for both MLP and XGBoost across all evaluation metrics, with R^2 = 0.997 for the former and R^2 = 0.918 for the latter. In contrast, linear regression showed only marginal improvement, reinforcing that its limitations stem from its inherent linear formulation rather than from the presence of irrelevant features.

Based on cross-validation results, the MLP trained on the reduced dataset was selected as the final model due to its superior predictive performance and subsequently evaluated on the independent test set. The results were consistent with those obtained during cross-validation, indicating strong generalization capability. A summary of the models performances across all stages, including cross-validation metrics and test set results, is presented in Table 1. Overall, the findings demonstrate that a compact set of key economic and process variables, combined with a relatively simple neural network architecture, is sufficient to accurately predict the economic viability of the process. This approach provides a computationally efficient alternative to detailed process simulations for preliminary profitability assessments.

Table 1: Performance comparison of machine learning models for IRR prediction.

Metrics	**Cross-validation results (average of validation sets)**						
	Full dataset			Reduced dataset			Test set
	Linear regression	XGBoost	MLP	Linear regression	XGBoost	MLP	MLP
R^2	0.172	0.896	0.996	0.176	0.918	0.997	0.998
RMSE	2.393	0.849	0.167	2.388	0.754	0.136	0.122
MAE	1.276	0.222	0.042	1.271	0.196	0.036	0.032

3.2 SHAP analysis

The SHAP analysis of the MLP model reveals that ethanol price is the most influential variable affecting IRR, highlighting the importance of the ethanol and bioethanol market value to the economic viability of the proposed plant. The prices of benzene and higher alcohol products also significantly affect IRR predictions, while ethylene glycol does not appear as a dominant economic driver, mainly due to lower flowrate and higher separation efficiency.

Financial parameters, such as federal income tax rate and, in particular, equity contribution, also exert a strong influence in the model predictions, suggesting that the capacity to sustain higher upfront capital investments is a key determinant in the investment decision. Capital cost parameters, however, are not among the primary drivers of the plant profitability, with the exception of the heat exchangers cost, which is consistent with the energy-intensive nature of the process. Overall, the SHAP analysis confirms that economic variables directly linked to feedstock costs and financial parameters dominate the model's predictions, while capital cost parameters and product revenues play a secondary, yet still relevant, role in determining the plant's economic feasibility.

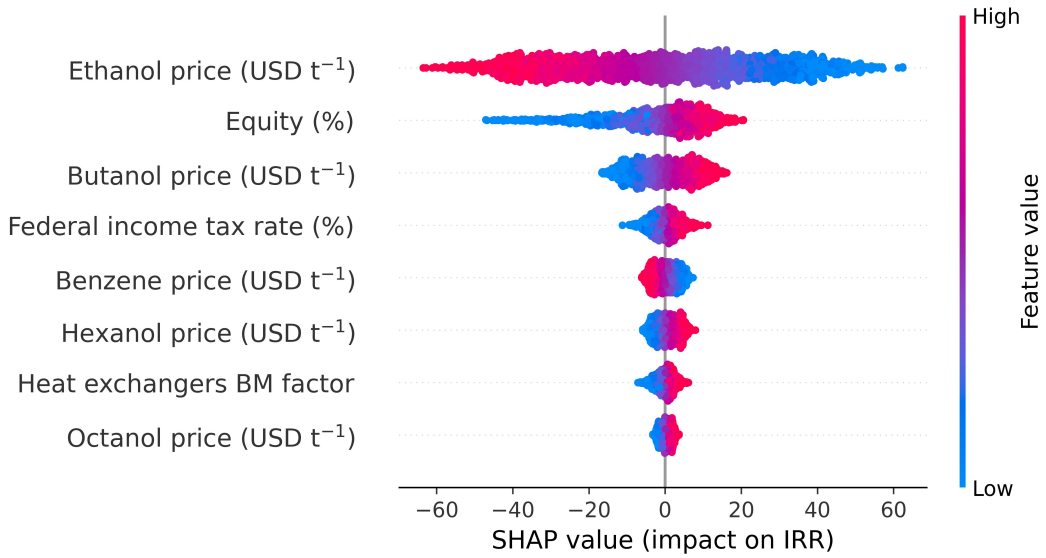

Figure 2: Effect of the most important features according to SHAP values.

4 Conclusions

The present work highlights the potential of data-driven surrogate models for predicting the economic performance of Guerbet coupling processes. Among the evaluated approaches, the MLP model showed superior performance, achieving high accuracy and strong generalization capability, particularly after feature selection. The results highlight that the nonlinear relationships inherent to the process cannot be adequately captured by linear models, reinforcing the importance of using more flexible machine learning techniques for techno-economic analysis. Furthermore, the reduction of the feature space improved model performance, indicating that a limited set of key variables is sufficient to describe the system behavior with high fidelity.

From an economic perspective, the SHAP analysis provided valuable insights into the factors driving process profitability. The ethanol price was identified as the most critical variable, exerting a strong negative influence on IRR when increased, while product prices, especially butanol, positively impacted economic viability. Additionally, financial parameters such as the equity fraction were shown to play a relevant role, highlighting the importance of capital structure in project feasibility.

The integration of process simulation and machine learning proved to be a powerful strategy for accelerating techno-economic evaluations. By significantly reducing computational time while maintaining high predictive accuracy, the proposed approach enables rapid screening of process scenarios and supports decision making in early stage design. Future research is expected to focus on improving the separation system and exploring alternative strategies that avoid the use of benzene as an entrainer, given its toxicity.

Acknowledgements: The present work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

Bušić, A., Mardetko, N., Kundas, S., Morzak, G., Belskaya, H., Šantek, M. I., Komes, D., Novak, S., & Šantek, B. (2018). Bioethanol production from renewable raw materials and its separation and purification: A review. https://doi.org/10.17113/ftb.56.03.18.5546

Chen, T., & Guestrin, C. (2016). Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Gabriëls, D., Hernández, W. Y., Sels, B., Voort, P. V. D., & Verberckmoes, A. (2015). Review of catalytic systems and thermodynamics for the guerbet condensation reaction and challenges for biomass valorization. *Catalysis Science & Technology*, *5*, 3876–3902. https://doi.org/10.1039/C5CY00359H

Hamden, Z., El-Ghoul, Y., Alminderej, F. M., & Majdoub, H. (2025). Biomass valorization toward sustainable bioethanol production: A critical review of feedstocks and techno-economic aspects. https://doi.org/10.1007/s12649-025-03339-4

Jin, C., Yao, M., Liu, H., Lee, C.-f. F., & Ji, J. (2011). Progress in the production and application of n-butanol as a biofuel. *Renewable and Sustainable Energy Reviews*, *15*, 4080–4106. https://doi.org/10.1016/j.rser.2011.06.001

Kozlowski, J. T., & Davis, R. J. (2013). Heterogeneous catalysts for the guerbet coupling of alcohols. *ACS Catalysis*, *3*, 1588–1600. https://doi.org/10.1021/cs400292f

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777. https://doi.org/10.48550/arXiv.1705.07874

Ordonez, D. A. R., Mesquita, L. C. S., & Brandão, A. L. T. (2026). Machine learning-driven techno-economic uncertainty analysis in batch pharmaceutical manufacturing. *ACS Omega*. https://doi.org/10.1021/acsomega.5c09455

Rocha, R., Colchete, A., Netto, L. D. S., Silva, M. C. B., & Brandão, A. L. T. (2019). Design, optimization and economic evaluation for the production of higher alcohols from bioethanol. *I Process Systems Engineering - PSE-BR2019*. https://doi.org/10.13140/RG.2.2.22751.82081

Seider, W. D., Lewin, D. R., & Widagdo, S. (2003). *Product and process design principles: Synthesis, analysis and evaluation* (2nd ed.).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tsuchida, T., Sakuma, S., Takeguchi, T., & Ueda, W. (2006). Direct synthesis of n-butanol from ethanol over nonstoichiometric hydroxyapatite. *Industrial and Engineering Chemistry Research*, *45*, 8634–8642. https://doi.org/10.1021/ie0606082

Turton, R., Bailie, R. C., Whiting, W. B., & Shaeiwitz, J. A. (2009). *Analysis, synthesis and design of chemical processes* (3rd ed.).

Zhang, Q., Dong, J., Liu, Y., Wang, Y., & Cao, Y. (2016). Towards a green bulk-scale biobutanol from bioethanol upgrading. *Journal of Energy Chemistry*, *25*, 907–910. https://doi.org/10.1016/j.jechem.2016.08.010