# Integration of Process Simulation and Machine Learning for Techno-Economic Evaluation of $CO_2$-Assisted Propane Dehydrogenation

Letícia Costa da Silva Mesquita[a], Guilherme Vieira Espinosa[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 present study investigates an integrated framework combining process simulation and machine learning to evaluate the profitability of propane dehydrogenation systems using $CO_2$ as a mild oxidant. Simulations were conducted using Aspen Plus to generate a comprehensive dataset covering economic, operational, and process variables. XGBoost (eXtreme Gradient Boosting) regression model was applied to predict $CO_2$ break-even price (BEP), achieving high accuracy and strong generalization performance. Feature selection using least absolute shrinkage and selection operator (LASSO) and model-based importance reduced the amount of features while improving both predictive accuracy and computational efficiency. SHAP (SHapley Additive exPlanations) analysis provided interpretability, revealing that economic parameters, such as equity, minimum acceptable rate of return, and federal income tax, are the primary drivers of BEP, while process variables showed limited influence within the studied range. The proposed approach highlights the potential of integrating process simulation tools with machine learning models to enhance techno-economic evaluations, enabling faster and more efficient decision-making in the optimization of chemical processes.

**Keywords:** Propylene production, Process simulation, Techno-economic evaluation, Machine learning

## 1 Introduction

Propylene, a key precursor to high-value chemicals such as polypropylene, propylene oxide, acrylonitrile, acetone and acrolein (Hu et al., 2019), has experienced a widening gap between supply and demand in recent years. In addition to traditional petrochemical processes (fluid catalytic cracking of heavy oil and steam cracking of naphta) and conventional catalytic propane dehydrogenation (M. L. Sun et al., 2023), the use of $CO_2$ as a mild oxidant suggests itself due to high availability and environmental benefits associated with $CO_2$ utilization. The increasing interest in $CO_2$-assisted propane dehydrogenation is aligned with broader global efforts to mitigate greenhouse gas emissions while producing high-value chemicals. Carbon capture, utilization, and storage (CCUS) technologies have been recognized as essential pathways toward achieving net-zero emission targets, particularly by enabling the conversion of $CO_2$ into valuable products such as methanol and light olefins (Symonds et al., 2025; Yang et al., 2024). In this context, the integration of $CO_2$ into reaction systems not only contributes to emission reduction but also enhances process sustainability.

To address these challenges, process simulation and advanced optimization techniques have become fundamental tools in the analysis of complex chemical systems. Integrated processes involving $CO_2$ capture and conversion typically exhibit strong nonlinear interactions among operating variables, making conventional optimization approaches computationally demanding and inefficient (Yang et al., 2024; Zhang et al., 2024). In recent years, machine learning (ML) methods have emerged as powerful alternatives, capable of accurately capturing these nonlinear relationships and providing fast predictions of key performance indicators. These models enable rapid screening of operating conditions and facilitate techno-economic and environmental assessments, significantly accelerating process design and decision-making. Furthermore, data-driven approaches based on machine learning have demonstrated strong potential to reduce dependence on extensive simulations and experimental campaigns. By training models on datasets generated from process simulators or industrial operations, it is possible to achieve high predictive accuracy while substantially lowering computational costs (Cunningham et al., 2025).

In this context, the present work aims to integrate Aspen Plus-based process simulations with machine

learning techniques to evaluate the profitability of an integrated oxidative dehydrogenation of propane using $CO_2$ (ODPC) and chemical looping combustion of propane to generate the reactor heat demand. By leveraging data generated from process simulation and applying regression-based ML model, this study seeks to capture the complex relationships between operating variables and economic performance, enabling the efficient evaluation of multiple operating scenarios and supporting more informed decision-making regarding process optimization and decarbonization strategies.

## 2 Methodology

### *2.1 Process description and dataset generation*

The proposed ODPC process design, described in greater detail by Espinosa and Brandão (2025), integrates a propane chemical looping combustion (CLC) and heat recovery section to the ODPC production to supply the dehydrogenation reactor's heat demand and electricity for the plant, pursuing energy self-sufficiency, and a carbon capture section (CCS). A simplified flow diagram of the plant is shown in Figure 1.

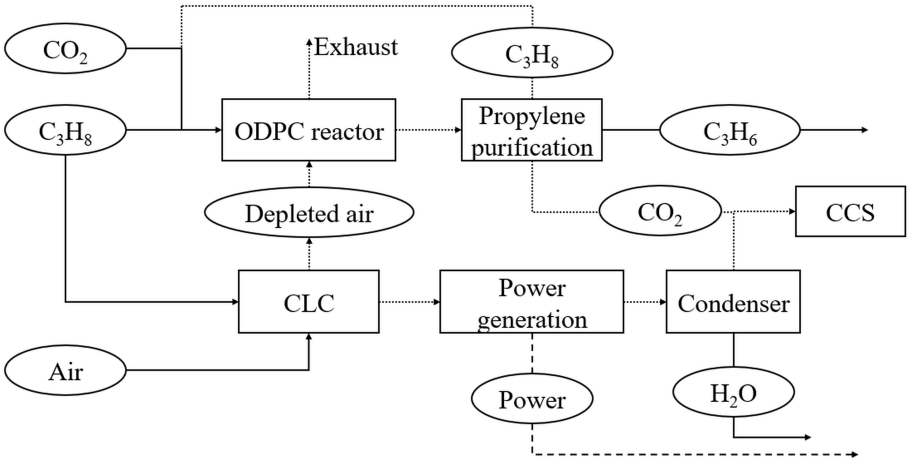


**Figure 1:** Simplified ODPC flow diagram.

The plant equipment costs were estimated using cost correlations described by Seider et al. (2003), with the exception of the CLC reactors which used Equation (1) for the air reactor and Equation (2) for the fuel reactor (Z. Sun et al., 2022).

$$C_{BM} = 0.69 \left(\frac{S}{59.4}\right)^{0.67} \left(\frac{CEPCI_p}{541.7}\right) \tag{1}$$

$$C_{BM} = 2.15 \left(\frac{S}{280.5}\right)^{0.67} \left(\frac{CEPCI_p}{541.7}\right) \tag{2}$$

Where $C_{BM}$ is the bare module cost, $CEPCI_P$ is the present year CEPCI (Chemical Engineering Plant Cost Index) and $S$ is the reactor capacity, which is the air mass flow ($kg\,s^{-1}$) for the air reactor and the oxygen carrier mass flow ($kg\,s^{-1}$) for the fuel reactor.

The operating costs were estimated using the multiplying factors described by Turton et al. (2009) where direct manufacturing costs, fixed manufacturing costs and general manufacturing costs are calculated using costs of raw material, waste treatment, utilities and labor as well as fixed investment cost. The ODPC plant profitability was evaluated using $CO_2$ break-even price (BEP) as performance metric, calculated with Equation (3).

$$NPV(BEP) = \sum_i \frac{CF_i(BEP)}{(1+MARR)^i} = 0 \tag{3}$$

Where $NPV$ is the net present value of the plant, $MARR$ is the minimum acceptable rate of return and $CF_i$ is the cash flow in year $i$.

The dataset generation was performed by varying parameters relating to costs composing the total investment cost and the cost of manufacturing, financial parameters such as operation time, federal income tax rate and MARR, process parameters such as plant capacity and reactor space velocity and temperature and prices of streams and utilities. Overall, 5,000 different combinations of process conditions and economic parameters were evaluated to generate the dataset.

### 2.2 *Machine learning workflow*

The $CO_2$ break-even price prediction task was formulated as a regression problem using the simulated dataset and the analysis was implemented in Python. The dataset was split into 85% for training and 15% for testing. The training subset was used in the cross-validation stage through the K-Fold technique, with five folds and random shuffling of the data. This procedure ensures that the model is exposed to different subsets during training, reducing the risk of bias associated with a single data split and promoting a more robust evaluation.

To ensure the reproducibility of the results, a random seed was fixed both in the data splitting stage and in the cross-validation process. Additionally, in each iteration of the validation process, normalization parameters (minimum and maximum values) were calculated exclusively from the training subset and then applied to both the training and validation sets. Consequently, all variables were rescaled to the range [0, 1]. This preprocessing step is particularly relevant for algorithms that are sensitive to the scale of input variables.

For the regression task, the eXtreme Gradient Boosting (XGBoost) model was adopted. In this study, the algorithm was used with its default hyper-parameters, aiming to evaluate its generalization capability even without specific tuning, especially when applied to well-structured datasets. The main performance metric adopted was the mean absolute percentage error (MAPE), as defined in Equation (4), while complementary metrics root mean square error (RMSE) and coefficient of determination ($R^2$) were also considered for a more comprehensive evaluation of the model, as presented in Equations (5) and (6).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{4}$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

Where $n$ is the number of samples, $y_i$ and $\hat{y}_i$ represent the true and predicted values of the $i$th data, respectively, and $\bar{y}$ denotes the average of the dataset on which the metric is being evaluated.

To reduce model complexity, a feature selection step was carried out by combining two complementary approaches: LASSO (least absolute shrinkage and selection operator) regression and the feature importance analysis provided by the XGBoost library. LASSO regression promotes coefficient penalization through L1 regularization, shrinking less relevant variables to zero and so yielding a sparse and more interpretable model, primarily capturing linear relationships (Tibshirani, 1996). In contrast, the feature importance metrics derived from XGBoost are based on each variable's contribution to the reduction of the loss function across decision tree splits, allowing the identification of nonlinear effects and complex interactions among features (Chen & Guestrin, 2016). The combination of these methods provides a more robust feature selection framework by evaluating variable relevance from different perspectives. Using the selected subsets of variables, the previously outlined training and validation process was carried out again, this time with a reduced feature set. The model that achieved the best performance was then assessed on the test dataset.

In order to enhance the interpretability of the results, an analysis based on SHAP (SHapley Additive exPlanations) was performed, grounded in cooperative game theory. This method assigns an importance value to each feature by quantifying its contribution to a specific prediction, considering different feature combinations. SHAP represents the model output as an additive sum of feature attributions that satisfies properties such as local accuracy and consistency, ensuring a theoretically sound interpretation of feature influence on the target variable (Lundberg & Lee, 2017).

## 3 Results and discussion

### *3.1 Prediction of $CO_2$ break-even price with XGBoost*

In the first stage of the analysis, the $CO_2$ break-even price was estimated using the full set of 41 variables from the simulated dataset. The total computational time required for this step was 6.62 seconds. Even when considering a relatively large set of input features, the model demonstrated promising predictive performance with $R^2 = 0.966$, indicating that the XGBoost algorithm is capable of effectively capturing the underlying relationships in the data.

When applying the previously described feature selection strategy, both the LASSO regression and XGBoost feature importance scores identified the same subset of key variables: MARR, equity contribution, capacity, propane price, space velocity, electricity price, federal income tax rate and the costs of compressors and heat exchangers, represented by their bare module factors. The agreement between both approaches in identifying the same subset of key variables reinforces the reliability of the selected features, suggesting that they have a consistent and significant impact on the predictive performance of the model and, based on these findings, the dataset was reduced to include only the nine selected variables.

Although the model trained on the full dataset already demonstrated robust performance, the reduced feature set resulted in improvements across all validation metrics, while also decreasing the computational time to 1.34 seconds. These results indicate that feature selection not only enhances predictive accuracy but also improves computational efficiency. After the cross-validation procedure using the reduced feature set, the final model was evaluated on an independent test dataset to assess its generalization performance on unseen data. The results of the cross-validation step, expressed in terms of $R^2$, RMSE and MAPE, as well as the simplified model performance on the test set, are summarized in Table 1.

**Table 1:** Performance of the model in different scenarios.

| | **Cross-validation results (average)** | | **Test set** |
|---|---|---|---|
| **Metrics** | **Full dataset** | **Reduced dataset** | **Reduced dataset** |
| $R^2$ | 0.966 | 0.970 | 0.971 |
| RMSE | 3.811 | 3.567 | 3.370 |
| MAPE (%) | 0.011 | 0.010 | 0.941 |

The results on the test dataset indicate that the model maintained high predictive performance and demonstrated a strong capability to generalize to unseen data, with the consistency between cross-validation and test set performance reinforcing the consistency of the proposed approach. The combination of feature selection and the XGBoost model proved effective in delivering accurate and reliable $CO_2$ break-even price predictions, demonstrating its potential as a robust tool to evaluate profitability of the process plant.

### *3.2 SHAP analysis*

The SHAP method was employed to interpret the predictions of the XGBoost model and quantify the contribution of each input variable to the $CO_2$ break-even price prediction. This approach provide interpretability by assigning an importance value to each feature based on its impact on the model output. In this study, SHAP summary, as shown in Figure 2, was used to evaluate not only the relative importance of the variables but also the direction of their influence across all observations.

The plot indicates that financial parameters play a dominant role in the model prediction, as MARR, equity contribution and federal income tax rate exhibit large spreads of SHAP values. As expected, increases in MARR and federal income tax rate lead to higher $CO_2$ break-even prices, as both factors directly reduce the net present value (Equation (3)). Similarly, a higher equity contribution adversely affects plant profitability, which can be attributed to the capital-intensive nature of the plant. Conversely, larger plant capacities enhance the project's economic performance, as expected from the usual scaling behavior described by equipment cost correlations (Seider et al., 2003).

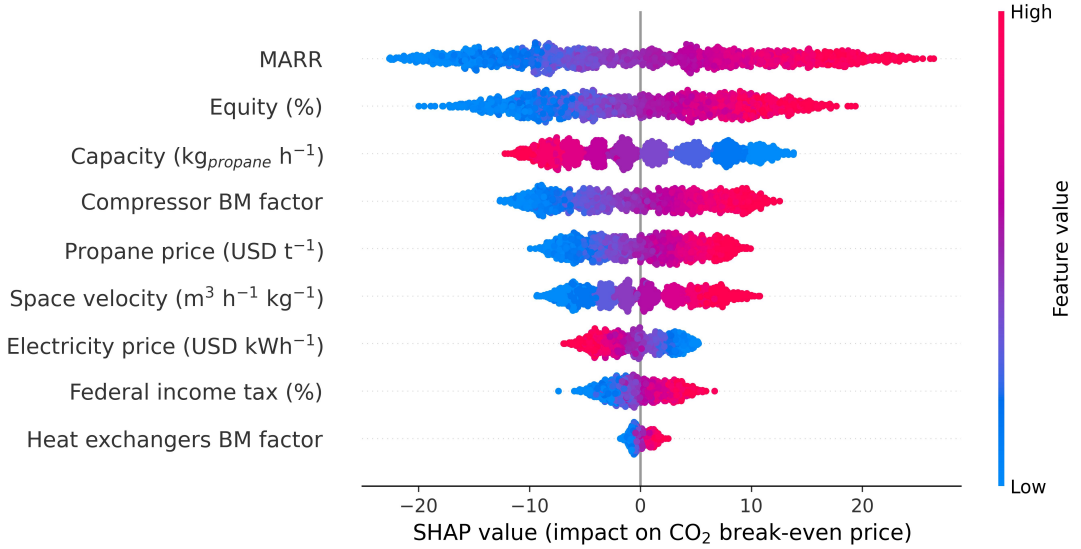


**Figure 2:** Behavior of the most influential variables according to SHAP values.

The SHAP plot further reveals the project's investment cost is largely influenced by the carbon capture section, with the cost contribution of compressors acting as a key driver to the model break-even price predictions. While operating costs do not appear to influence the prediction behavior, propane price, in particular, exerts a strong impact on the project's performance. In contrast, revenues derived from propylene contribute marginally to the plant's profitability as it is outweighed by the larger revenues stemmed from the electricity generated in the heat recovery section, as well as carbon credit revenues.

These results demonstrate that market conditions exert a predominant influence on the project's economic performance, directly determining the magnitude of carbon abatement incentives require to guarantee economic viability. Additionally, prospective technological advancements of carbon capture technologies and the associated reductions in equipment cost may substantially improve the profitability of the ODPC process. Regarding decision variables, the SHAP analysis demonstrates that larger capacities and lower ODPC reactor space velocities reduce the $CO_2$ break-even price, with increasing equity contribution also exerting a favorable effect. As such, the capacity to sustain higher upfront capital investment is a key determinant of both the profitability and the carbon abatement incentive requirements of the proposed ODPC plant.

## 4 Conclusions

This study demonstrated the effectiveness of integrating process simulation with machine learning techniques for the prediction of $CO_2$ break-even price in the proposed integrated ODPC and carbon capture system. The XGBoost model exhibited robust predictive performance, achieving high accuracy and strong generalization capability, as evidenced by the $R^2$ values of 0.9666 and 0.9692 in both cross-validation and independent test datasets. Furthermore, the application of feature selection techniques enabled a reduction in problem dimensionality, resulting in improved computational efficiency and enhanced predictive performance. These findings highlight the relevance of combining statistical learning methods with process-based modeling in the analysis of complex systems. Additionally, the SHAP analysis provided interpretability to the model predictions, identifying financial parameters, plant capacity, ODPC reactor space velocity, propane and electricity prices and capital investment in the carbon capture section as the primary drivers of the $CO_2$ break-even price. As such, the present work demonstrates that the decision to invest is primarily governed by the capacity to sustain higher upfront capital investment.

The proposed framework demonstrates the potential of combining machine learning with conventional process simulation tools as a powerful strategy for the analysis of complex chemical systems. This integrated approach enables fast and accurate prediction of key performance indicators, reducing reliance on computationally intensive

simulations and facilitating the exploration of the techno-economic evaluation process.

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

**References**

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

Cunningham, R. D., Patterson, V., Cawood, E., Botes, G., & Marantos, E. (2025). Data-driven optimization of industrial impact polypropylene characterization: Machine learning insights. *Journal of Chemical Information and Modeling*, *65*, 6655–6670. https://doi.org/10.1021/acs.jcim.5c00667

Espinosa, G. V., & Brandão, A. L. (2025). Economic and sustainability evaluation of green co2-assisted propane dehydrogenation design. *Digital Chemical Engineering*, *14*, 100203. https://doi.org/10.1016/j.dche.2024.100203

Hu, Z.-P., Yang, D., Wang, Z., & Yuan, Z.-Y. (2019). State-of-the-art catalysts for direct dehydrogenation of propane to propylene. *Chinese Journal of Catalysis*, *40*, 1233–1254. https://doi.org/10.1016/S1872-2067(19)63360-7

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

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

Sun, M. L., Hu, Z. P., Wang, H. Y., Suo, Y. J., & Yuan, Z. Y. (2023). Design strategies of stable catalysts for propane dehydrogenation to propylene. *ACS Catalysis*, *13*, 4719–4741. https://doi.org/10.1021/acscatal.3c00103

Sun, Z., Wang, S., & Aziz, M. (2022). Highly integrated system for ammonia and electricity production from biomass employing direct chemical looping: Exergy and exergoeconomic analyses. *Energy Conversion and Management*, *251*, 115013. https://doi.org/10.1016/j.enconman.2021.115013

Symonds, R., Shokrollahi, M., Hughes, R., Navarri, P., & Modler, R. (2025). Systematic approach to the design, modeling, and techno-economic-environmental analysis of co2 capture technologies as part of the national ccus assessment framework (ncaf). *Carbon Capture Science and Technology*, *16*. https://doi.org/10.1016/j.ccst.2025.100439

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

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

Yang, Q., Zhou, J., Bao, R., Rong, D., & Zhang, D. (2024). Machine learning-based techno-econo-environmental analysis of co2-to-olefins process for screening the optimal catalyst and hydrogen color. *Energy*, *312*. https://doi.org/10.1016/j.energy.2024.133508

Zhang, Z., Vo, D. N., Nguyen, T. B., Sun, J., & Lee, C. H. (2024). Advanced process integration and machine learning-based optimization to enhance techno-economic-environmental performance of co2 capture and conversion to methanol. *Energy*, *293*. https://doi.org/10.1016/j.energy.2024.130758