MLP Surrogate Modeling with SHAP Interpretability for CO_2 Absorption in Packed Columns via AVEVA Process Simulation

Vinicius Seferian Scheffer Machado[a], Carlos Alexandre Moreira da Silva[a], Nicolas Spogis[b], Milene Costa Codolo[a*]

[a]UNIFESP, Institute of Environmental, Chemical and Pharmaceutical Sciences, Diadema-SP, Brazil
[b]UNICAMP, School of Chemical Engineering, Campinas-SP, Brazil

*milene.codolo@unifesp.br

ABSTRACT

This work presents a multi-layer perceptron (MLP) surrogate model with SHAP interpretability for a CO_2 chemical absorption process in a packed column using NaOH as absorbent, simulated in AVEVA Process Simulation (APS). A dataset of 4,000 steady-state scenarios was generated via Latin Hypercube Sampling across five input variables: gas flow rate (10-100 m^3/h), NaOH feed rate (100-1,000 L/h), column height (1-5 m), column diameter (0.5-1.0 m), and CO_2 inlet concentration (0.2-1.5 vol.%). The MLP architecture, optimized through Optuna Bayesian search (100 trials), simultaneously predicts five process outputs: removal efficiency, gas-phase and liquid-phase mass transfer coefficients, column capacity factor, and interfacial area. The surrogate achieved a mean R^2 of 0.9992 across all outputs with MAPE below 2.72%. SHAP analysis revealed that gas flow rate dominates removal efficiency (normalized importance = 1.00), while column diameter governs mass transfer coefficients (0.77-0.82). CO_2 inlet concentration showed negligible global impact (importance = 1.07), confirming findings from classical DOE analysis. The surrogate enables real-time process exploration at negligible computational cost compared to the rigorous rate-based APS simulation.

Keywords: CO_2 absorption; packed column; surrogate model; SHAP; AVEVA Process Simulation.

1. Introduction

Carbon dioxide emissions from fossil fuel combustion remain a critical environmental challenge, driving the development of carbon capture, utilization, and storage (CCUS) technologies. Among the available post-combustion capture methods, chemical absorption in packed columns is the most mature and widely deployed technology (Rochelle, 2009). The design and optimization of these systems involve complex interactions between fluid dynamics, mass transfer kinetics, and chemical reaction, requiring rigorous process simulation tools.

Machado (2024) presented a comprehensive steady-state simulation study of CO_2 absorption in a packed column using NaOH as absorbent in AVEVA Process Simulation (APS), employing a rate-based column model with the e-NRTL thermodynamic framework. That work investigated the influence of operating variables through classical 2^5 factorial design of experiments (DOE) with 35 scenarios, achieving linear model fits with R^2 above 0.99. However, the factorial approach is limited to linear and binary interaction effects, potentially missing nonlinear dependencies across the wide operating range explored.

Machine learning (ML) surrogate models have emerged as powerful tools for replacing computationally expensive process simulations with fast, accurate approximations. Multi-layer perceptrons (MLPs) are particularly suited for regression tasks in chemical engineering due to their universal approximation capability (Schweidtmann et al., 2021). When combined with SHapley Additive exPlanations (SHAP), these models provide not only predictive accuracy but also interpretable insights into feature importance and interaction effects (Lundberg and Lee, 2017).

This work extends the simulation framework of Machado (2024) by generating 4,000 scenarios via Latin Hypercube Sampling (LHS) across expanded operating ranges, training an Optuna-optimized MLP surrogate that simultaneously predicts five process outputs, and applying SHAP analysis to provide output-specific interpretability that goes beyond classical DOE. The complete code and dataset are available at https://github.com/Spogis/ml-shap-co2-absorption.

2. Methodology

2.1. Process simulation and dataset generation

The process model consists of a rate-based packed absorption column simulated in AVEVA Process Simulation 2024.1, using the e-NRTL thermodynamic model for the CO_2-NaOH-H_2O electrolyte system. The column operates in counter-current mode with NaOH solution fed at the top and the gas mixture (air + CO_2) entering at the bottom. Five transfer units with 100% contact efficiency were adopted as established by preliminary convergence studies (Machado, 2024).

A dataset of 4,000 steady-state scenarios was generated using Latin Hypercube Sampling across five input variables: gas flow rate Q_G (10-100 m^3/h), NaOH feed rate Q_L (100-1,000 L/h), column height h_{col} (1.0-5.0 m), column diameter d_{col} (0.5-1.0 m), and CO_2 inlet concentration y_{CO2} (0.2-1.5 vol.%). The simulation outputs comprise five response variables: CO_2 removal efficiency, gas-phase mass transfer coefficient k_G ($kmol/m^2.s.atm$), liquid-phase mass transfer coefficient K_L (m/s), column capacity factor F_{col}, and effective interfacial area A_{it} (m^2/m^3). Table 1 summarizes the variable ranges.

Table 1. Input and output variable ranges for the 4,000-scenario dataset.

Variable	Role	Min	Max	Unit
Gas Flow Rate (Q_G)	Input	10	100	m^3/h
NaOH Feed Rate (Q_L)	Input	100	1,000	L/h
Column Height (h_{col})	Input	1.0	5.0	m
Column Diameter (d_{col})	Input	0.5	1.0	m
CO_2 Concentration (y_{CO2})	Input	0.2	1.5	vol.%
Removal Efficiency	Output	2.80	88.75	%
k_G	Output	0.0012	0.0150	$kmol/m^2.s.atm$
K_L	Output	2.96e-5	1.65e-4	m/s
F_{col}	Output	0.0039	0.1499	-
A_{it}	Output	15.88	380.02	m^2/m^3

2.2. MLP surrogate model

The surrogate model employs a multi-layer perceptron (MLP) implemented in Keras/TensorFlow. The dataset was split into 80% training (3,200 samples) and 20% test (800 samples), with both inputs and outputs standardized via StandardScaler. Hyperparameter optimization was performed using Optuna Bayesian search with the Tree-structured Parzen Estimator (TPE) sampler over 100 trials, exploring the search space defined in Table 2.

Table 2. Optuna hyperparameter search space and selected values.

Hyperparameter	Search Space	Selected
Hidden layers	1 - 4	3
Neurons per layer	{32, 64, 128, 256}	256
Activation function	{ReLU, tanh, SELU}	ReLU
Learning rate	1e-5 - 5e-4 (log)	4.82e-4
Weight decay	1e-6 - 1e-3 (log)	5.00e-5
Batch size	Fixed	512
Max epochs	Fixed	2,000
Early stopping	patience = 50	val_loss

2.3. SHAP interpretability analysis

SHapley Additive exPlanations (SHAP) based on cooperative game theory were computed using KernelExplainer with 500 background samples and nsamples = 300 perturbations per prediction (Lundberg and Lee, 2017). The analysis generates per-output Shapley values that quantify each input variable contribution to individual predictions. Results are presented as: (i) a normalized heatmap where each output row is independently scaled to [0, 1], enabling cross-output comparison of relative feature importance; (ii) beeswarm plots showing the

distribution and direction of SHAP values; and (iii) a global importance bar chart averaging mean |SHAP| across all five outputs.

3. Results and Discussion

3.1. MLP surrogate performance

The optimized MLP architecture (5-256-256-256-256-5, ReLU activation) achieved excellent predictive performance across all five output variables. Table 3 presents the R^2 and MAPE metrics on the test set (800 unseen samples). The mean R^2 of 0.9992 indicates near-perfect generalization, with all individual outputs exceeding R^2 = 0.998. The gas-phase mass transfer coefficient k achieved the highest accuracy (R^2 = 0.9998, MAPE = 1.11%), while removal efficiency showed the largest relative error (MAPE = 2.71%) due to its wider dynamic range (2.80-88.75%) and stronger nonlinear dependence on multiple inputs.

Table 3. MLP surrogate model performance on the test set (800 samples).

Output Variable	R2	MAPE (%)
CO2 Removal Efficiency	0.9985	2.71
kG (gas-phase)	0.9998	1.11
KL (liquid-phase)	0.9986	0.96
Fcol (capacity factor)	0.9997	2.47
Ait (interfacial area)	0.9995	1.48
Mean	**0.9992**	**1.75**

Figure 1 presents the parity plots comparing simulated values from APS against MLP predictions for all five outputs. The tight clustering around the ideal diagonal confirms the model captures the full nonlinear behavior of the rate-based simulation across the entire operating space. No systematic bias was observed in any output, and the test set performance closely matches the training set, indicating no overfitting.

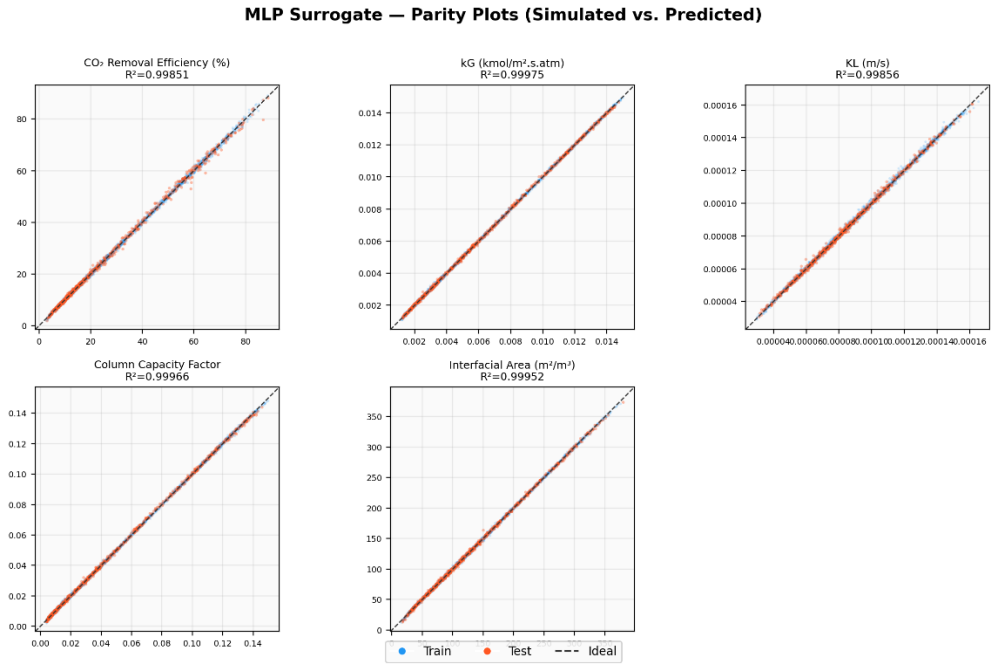

Figure 1. Parity plots (simulated vs. predicted) for all five outputs on train (blue) and test (red) sets.

3.2. SHAP interpretability analysis

Figure 2 presents the normalized SHAP heatmap, where each output row is independently scaled to [0, 1] to reveal relative feature importance patterns. The analysis reveals distinct input-output dependency structures that a classical linear DOE cannot fully capture:

For CO_2 removal efficiency, gas flow rate is the dominant driver (normalized importance = 1.00), followed by CO_2 concentration (0.36), NaOH feed rate (0.32), and column height (0.30). Column diameter has minimal influence (0.07). This hierarchy is consistent with the physical understanding: increasing gas velocity reduces residence time, directly decreasing removal. The SHAP beeswarm plots (Figure 3) confirm the negative directionality of gas flow, with high values (red) producing strongly negative SHAP contributions.

For the gas-phase mass transfer coefficient k_G, gas flow rate (1.00) and column diameter (0.77) are the only significant inputs, with all other variables showing negligible importance (0.01). This is physically consistent with the Onda et al. (1968) correlation used by APS, where k_G depends on the gas-phase Reynolds number, which is governed by superficial gas velocity (a function of Q_G and d_{col}). Notably, NaOH feed rate and column height have essentially zero impact on k_G, confirming the phase-specific decoupling of mass transfer coefficients.

The liquid-phase mass transfer coefficient K_L shows a complementary pattern: NaOH feed rate dominates (1.00), followed by column diameter (0.82), while gas flow contributes modestly (0.19). Again, this aligns with the Onda correlation, where K_L depends on the liquid-phase Reynolds number. The interfacial area A_{it} is primarily governed by column height (1.00) and diameter (0.77), with NaOH feed providing secondary influence (0.50).

Critically, CO_2 inlet concentration shows negligible SHAP importance across all outputs except removal efficiency (0.36). Its global importance (1.07) is the lowest among all inputs, confirming the finding of Machado (2024) that concentration effects in the 2,000-4,000 ppm range are statistically insignificant. The SHAP analysis extends this conclusion to the broader 0.2-1.5 vol.% range and provides output-specific granularity: while CO_2 concentration does influence efficiency (as expected from mass balance), it has virtually no effect on mass transfer coefficients, column capacity, or interfacial area.

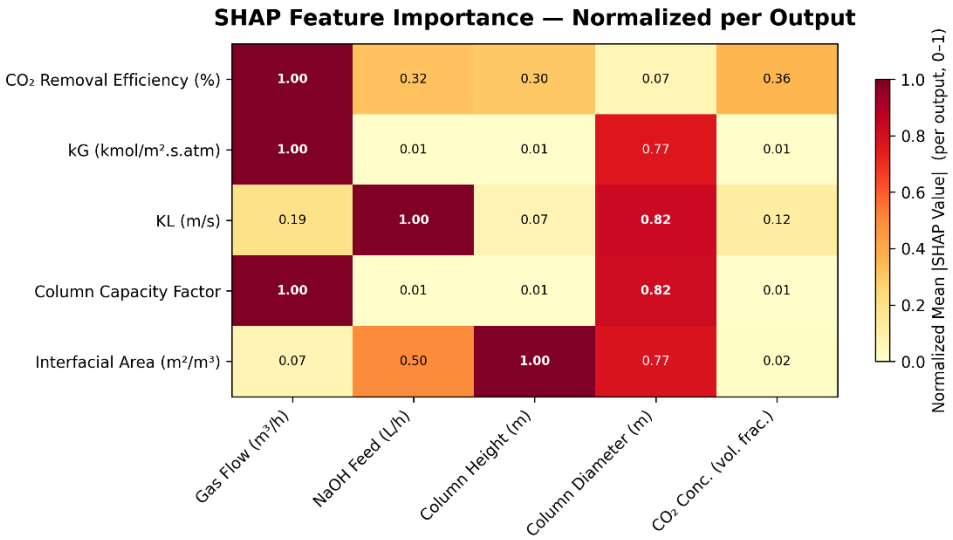

Figure 2. Normalized SHAP feature importance heatmap (per output, scaled 0-1).

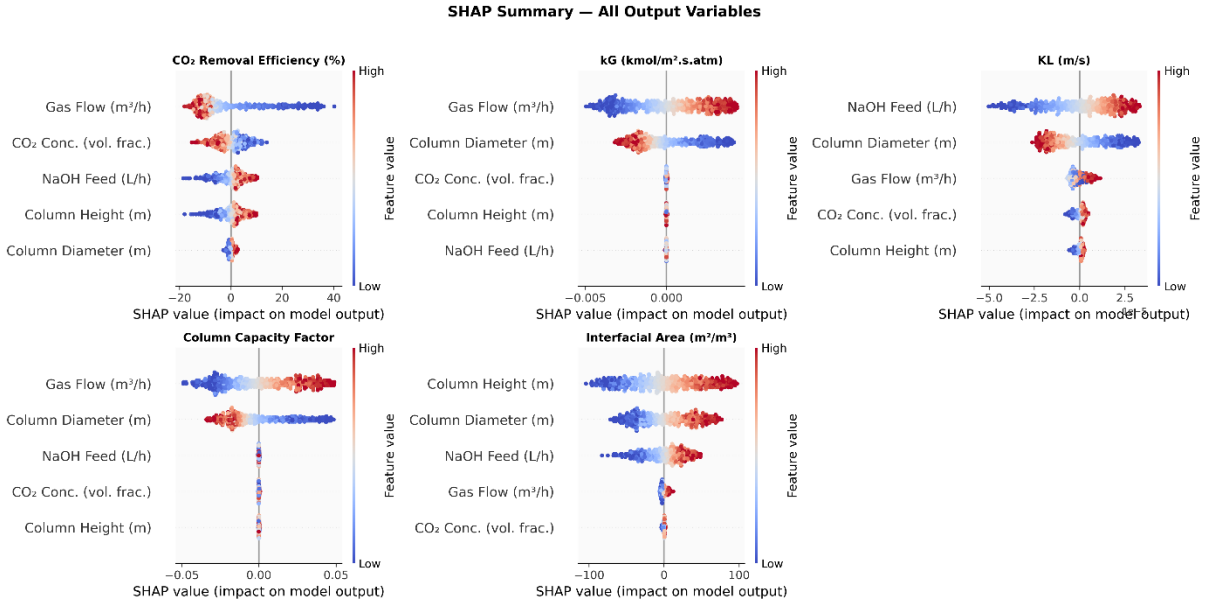

Figure 3. SHAP beeswarm summary plots for all five output variables.

4. Conclusions

An MLP surrogate model was developed for CO_2 absorption in packed columns, achieving mean R^2 = 0.9992 across five simultaneous outputs using 4,000 LHS-generated scenarios from AVEVA Process Simulation. The Optuna-optimized architecture (3 hidden layers, 256 neurons, ReLU) captures the full nonlinear behavior of the rate-based model with MAPE below 2.72% for all outputs.

SHAP analysis provided output-specific interpretability revealing that: (i) gas flow rate dominates removal efficiency with strong negative directionality; (ii) mass transfer coefficients exhibit phase-specific dependencies consistent with the Onda correlation, where k_G is controlled by gas velocity and diameter while KL is governed by liquid flow and diameter; (iii) CO_2 inlet concentration has negligible impact on mass transfer parameters, confirming and extending the classical DOE findings of Machado (2024) across a wider operating range; and (iv) column height is the most globally important variable (mean |SHAP| = 9.37), primarily through its influence on interfacial area and residence time.

The surrogate model enables real-time parametric exploration and can serve as the computational core for optimization, digital twin applications, and model predictive control of CO_2 absorption processes. The complete code and dataset are available at https://github.com/NicolasSpogworthy/ml-shap-co2-absorption.

Acknowledgments: The authors acknowledge UNIFESP for providing the computational infrastructure and AVEVA for the academic license of AVEVA Process Simulation.

References

S. M. Lundberg and S. I. Lee: A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems (30), 4766-4777, 2017.

V. S. S. Machado: Study of CO_2 Absorption in Packed Columns Using AVEVA Process Simulation, M.Sc. Dissertation, Universidade Federal de São Paulo, Diadema, 2024.

K. Onda, H. Takeuchi and Y. Okumoto: Mass Transfer Coefficients between Gas and Liquid Phases in Packed Columns, Journal of Chemical Engineering of Japan (1), 56-62, 1968.

G. T. Rochelle: Amine Scrubbing for CO2 Capture, Science (325), 1652-1654, 2009.

A. M. Schweidtmann, J. G. Rittig, A. König, M. Grohe, A. Mitsos and M. Dahmen: Graph Neural Networks for Prediction of Fuel Ignition Quality, Energy and Fuels (34), 11395-11407, 2020.