Development of a Machine Learning-Based Symbolic Regression Model for Mixing Time Prediction in Large Petroleum Storage Tanks

Reynaldo Pires da Fonseca[a,b], Diener Volpin Ribeiro Fontoura[a], Nicolas Spogis[a*], Dirceu Noriler[a], Guilherme José de Castilho[a], José Roberto Nunhez[a]

[a]*School of Chemical Engineering, University of Campinas (UNICAMP), Campinas-SP, Brazil*
[b]*Petrobras (Petróleo Brasileiro S.A.), Rio de Janeiro-RJ, Brazil*
**nicolas.spogis@gmail.com*

ABSTRACT

This study presents the development of a Symbolic Regression (SR) model, based on the PySR library, for predicting mixing time in large petroleum storage tanks equipped with side-entry impellers. A systematic methodology integrating computational fluid dynamics (CFD) simulations, factorial design of experiments (DOE), and statistical analysis (Spearman correlation and ANOVA) was employed to generate a comprehensive dataset covering 16 configurations. The key parameters investigated include oil viscosity, tank diameter, tank height, impeller tilt angle, and number of impellers. Spearman correlation and ANOVA revealed that the number of impellers is the dominant factor controlling mixing time ($\eta^2 = 0.534$, $p < 0.001$), followed by tank diameter and height. The SR model achieved a coefficient of determination $R^2 = 0.898$, yielding a compact, physically interpretable equation that captures the influence of geometric and operational variables on mixing performance. The proposed data-driven workflow provides an efficient tool for industrial-scale tank design and optimization.

Keywords: Symbolic Regression, Mixing Time, CFD, Side-Entry Impellers, Petroleum Storage Tanks.

1. Introduction

Large petroleum storage tanks equipped with side-entry impellers are critical assets in the oil and gas industry, where efficient mixing is essential to prevent sludge formation and ensure product homogeneity. Despite their industrial relevance, predictive models for mixing time in these large-scale systems remain scarce, particularly when compared to the extensive literature available for top-mounted impellers (Fonseca et al., 2025). The complex flow patterns generated by lateral impellers, characterized by tangential jets, toroidal recirculation zones, and stagnation regions, pose significant challenges for the development of reliable empirical correlations.

Computational Fluid Dynamics (CFD) has become an indispensable tool for analyzing mixing phenomena in stirred vessels. However, running full CFD simulations for every design configuration is computationally prohibitive, especially during early design stages when multiple geometric and operational combinations must be evaluated. This motivates the development of surrogate models that can rapidly predict mixing performance based on key design parameters.

Symbolic Regression (SR) is a machine learning technique that searches the space of mathematical expressions to find compact, interpretable equations that best fit the data (Cranmer, 2023). Unlike black-box methods such as neural networks, SR produces closed-form expressions with direct physical interpretability, making it particularly suitable for engineering applications. The PySR library, built on multi-population evolutionary algorithms, has demonstrated excellent performance in discovering physically meaningful correlations from simulation data.

The present work proposes an integrated methodology combining CFD simulations with factorial DOE, statistical analysis (Spearman correlation and ANOVA), and Symbolic Regression to develop a predictive model for mixing time in large petroleum storage tanks. The approach generates a compact, interpretable equation that captures the influence of five key parameters: oil viscosity, tank height, tank diameter, impeller tilt angle, and number of impellers.

2. Methodology

The computational domain consists of a vertical cylindrical storage tank with a flat bottom and a free surface (shear-free boundary). The three-dimensional geometry was constructed in SALOME and refined in Ansys DesignModeler, representing tanks with diameters ranging from 55 to 100 m, heights from 8 to 19 m, and 1 to 3 side-entry impellers. The impellers, based on actual dimensions provided by Petrobras, were incorporated through independent rotational domains using the Multiple Reference Frame (MRF) approach. Figure 1 shows the tank geometry and impeller positioning.

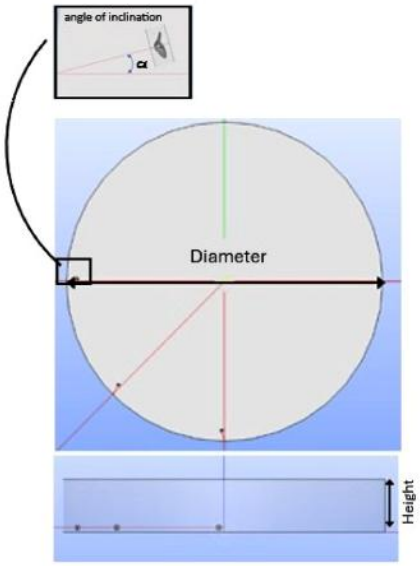

Figure 1 – Three-dimensional representation of the tank and impeller positioning – Fonseca et al., 2025.

Polyhedral meshes were generated individually in Ansys Meshing with localized refinement near the impellers and the opposite wall, where high velocity gradients occur. Prismatic inflation layers were applied at solid walls with $y+ \approx 1$, validated through boundary layer analysis. The total element count ranged from 2.5 to 4 million depending on tank dimensions. Grid independence was verified using the Grid Convergence Index (GCI) method, with discretization errors below 5% (Roache, 1994).

Steady-state simulations of the continuous phase were performed using ANSYS Fluent 2024 R2 with a pressure-based solver. The k-ω SST turbulence model (Menter, 1994) was employed due to its robustness in confined flows and near-wall recirculation zones. Second-order discretization was used for momentum and turbulent kinetic energy equations, with PRESTO pressure interpolation. The petroleum was modeled as an incompressible Newtonian fluid with density $\rho = 850$ kg/m³ and viscosity varying according to the DOE (10–55 cP). The impeller rotational speed was set at 410 rpm for all cases.

A fractional factorial DOE was applied to systematically evaluate the influence of five parameters on mixing time: viscosity (10–55 cP), tank height (8–19 m), tank diameter (55–100 m), impeller tilt angle (5°–20°), and number of impellers (1–3). A total of 16 independent CFD simulations were performed, each with its own geometry and mesh. The mixing time was determined from transient passive scalar simulations, where a miscible tracer was injected at the tank base. The flow was considered homogeneous when the relative variation of the mean concentration covariance between consecutive time steps fell below 1%, corresponding to 99% global homogeneity.

The relationship between mixing time and the DOE variables was assessed through Spearman rank correlation analysis, which identifies monotonic dependencies without assuming linearity. Analysis of Variance (ANOVA) was subsequently applied to quantify the statistical significance of each factor and its effect size (η^2). Based on the DOE results, a Symbolic Regression model was developed using the PySR library (Python), which employs multi-population evolutionary algorithms to discover compact mathematical expressions from data. The objective was to derive a simple, interpretable equation correlating geometric and operational variables to mixing time.

3. Results and Discussion

The CFD results revealed that the circulation pattern induced by side-entry impellers is characterized by a high-velocity tangential jet that propagates to the opposite wall, where it redirects downward, generating an intense mixing recirculation zone at the bottom. This toroidal flow pattern differs substantially from that observed in top-mounted impeller systems, where axial vortices and radial recirculation structures dominate. In the present configuration, the lateral flow generates a toroidal movement along the tank bottom, combining regions of high turbulent kinetic energy (k) near the jet with low-intensity stagnation zones susceptible to solid deposition. Figure 2 illustrates the velocity magnitude contour at the impeller centerline height.

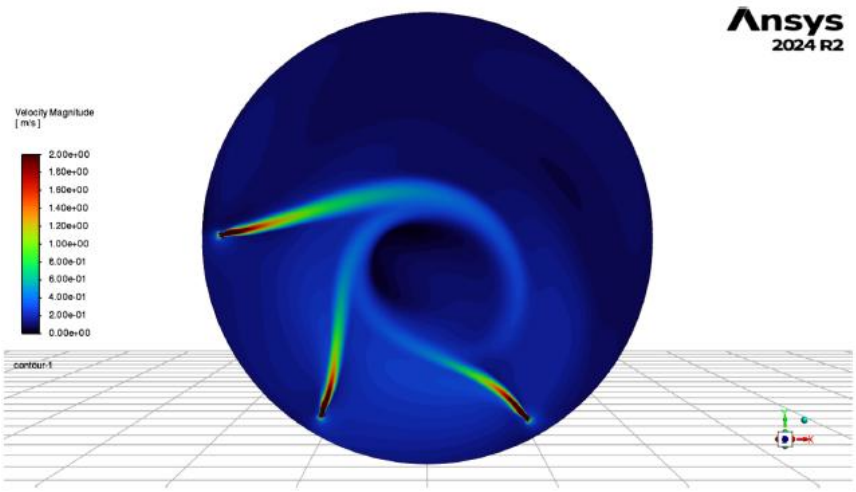

Figure 2 – Velocity field at the impeller centerline height inside the tank – Fonseca et al., 2025.

Increasing the number of impellers from one to two promoted a more symmetric flow distribution and uniform bottom coverage, significantly reducing low-velocity areas. With three impellers, the dead zone reduction was even more pronounced, with greater turbulent kinetic energy uniformity throughout the volume, evidencing the determinant role of this variable on global mixing efficiency.

The mixing time was computed for all 16 DOE configurations. The results demonstrated that the number of impellers is the most influential parameter: the configuration with three impellers reduced mixing time by more than 70% compared to a single impeller (from approximately 5–7 h down to 1.36 h). Tank diameter also proved determinant, with larger tanks requiring longer mixing times due to increased volume and recirculation distances. Viscosity had a secondary influence, raising mixing time by approximately 20% across the 10–55 cP range. The impeller tilt angle exhibited a non-linear response, with an optimal performance around 10°–15°, where balance between jet penetration and lateral sweep is achieved. Figure 3 shows a representative example of the temporal evolution of the passive scalar covariance used to determine the mixing time.

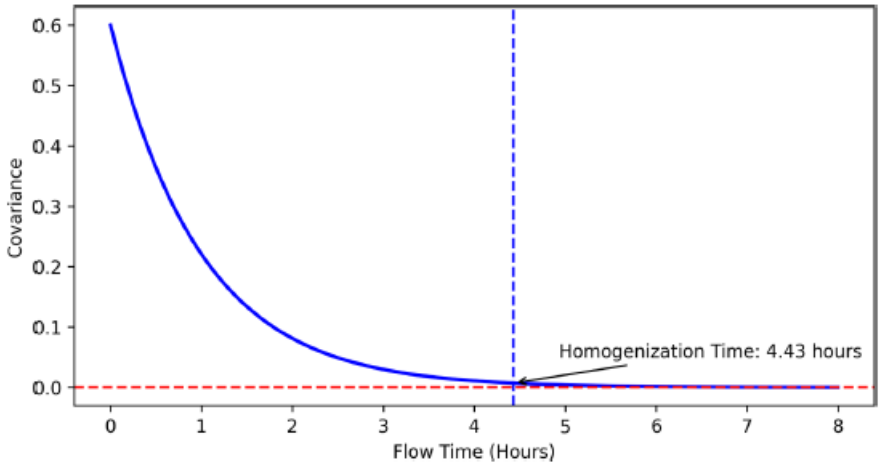

Figure 3 – Temporal evolution of the passive scalar concentration covariance (homogenization criterion) – Fonseca et al., 2025.

The Spearman rank correlation analysis (Figure 4) revealed a moderate positive correlation between mixing time and tank diameter ($\rho = 0.51$) and height ($\rho = 0.43$), indicating that larger tanks require longer homogenization periods. A weak positive correlation was observed with viscosity ($\rho = 0.23$), while the impeller tilt angle showed a weak negative correlation ($\rho = -0.15$). Most notably, the number of impellers exhibited a strong negative correlation ($\rho = -0.58$), confirming its role as the primary control variable for mixing performance.

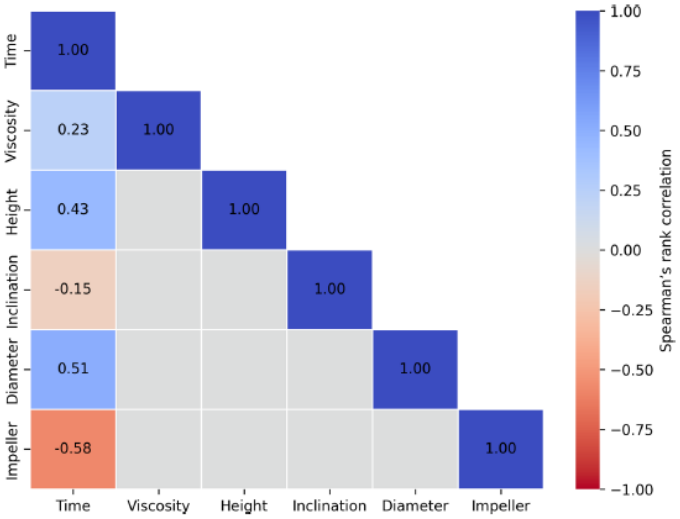

Figure 4 – Spearman rank correlation heatmap between mixing time and process variables – Fonseca et al., 2025.

The ANOVA results (Figure 5) provided quantitative confirmation. The number of impellers exhibited the largest standardized effect (8.237) and effect size ($\eta^2 = 0.534$, $p < 0.001$), followed by tank diameter ($\eta^2 = 0.291$, $p < 0.001$) and height ($\eta^2 = 0.145$, $p = 0.002$). Impeller tilt angle and viscosity did not reach statistical significance at the 5% level ($p = 0.171$ and $p = 0.239$, respectively), indicating marginal effects within the investigated ranges. These results confirm that mixing efficiency is dominated by geometric and agitation configuration factors, while fluid rheological properties play a secondary role.

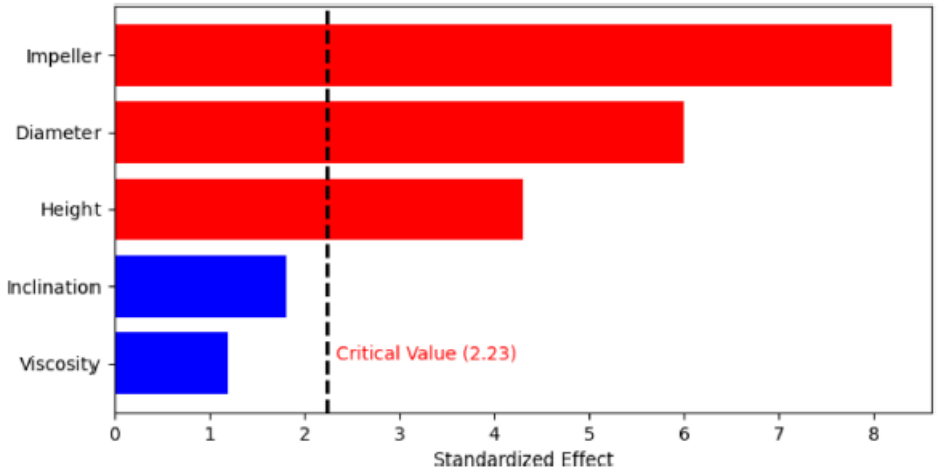

Figure 5 – ANOVA results: standardized effects on mixing time (critical value = 2.23) – Fonseca et al., 2025.

Using the PySR library, a Symbolic Regression model was fitted to the 16-point DOE dataset. The algorithm explored thousands of candidate mathematical expressions, optimizing the trade-off between complexity and accuracy through a Pareto front approach. The best model achieved $R^2 = 0.898$, representing an excellent fit to the CFD-simulated mixing time data. The resulting closed-form expression correlates mixing time with the geometric and operational variables in a compact, physically interpretable form (Fonseca et al., 2025).

The physical interpretation of the model is consistent with the statistical analysis: mixing time increases proportionally with tank diameter (D), exhibits a negligible dependence on height (H raised to the power of 0.007, effectively constant), and decreases with the number of impellers. This behavior reflects the dominant mechanisms identified through both Spearman correlation and ANOVA, confirming that the SR model captures the essential physics governing mixing in side-entry stirred tanks.

Compared to conventional empirical correlations that require a priori functional form assumptions, the SR approach offers a fully data-driven pathway to discover governing equations. The resulting expression can be directly employed for rapid screening of tank design configurations, eliminating the need for computationally expensive CFD simulations during preliminary engineering stages. This positions the methodology as a digital twin tool for industrial mixing process optimization.

4. Conclusions

An integrated methodology combining CFD simulations, Design of Experiments, statistical analysis, and Symbolic Regression was successfully developed for predicting mixing time in large petroleum storage tanks with side-entry impellers. The main conclusions are:

The number of impellers was identified as the dominant factor controlling mixing time ($\eta^2 = 0.534$), with three impellers reducing the homogenization time by over 70% compared to a single unit. Tank diameter and height also exhibited statistically significant positive effects, while viscosity and tilt angle showed marginal influence within the ranges investigated.

The Symbolic Regression model developed with PySR achieved $R^2 = 0.898$, yielding a compact, physically interpretable equation for mixing time prediction. The expression synthesizes the DOE results into a closed-form model suitable for rapid industrial-scale design evaluation, avoiding the computational cost of full CFD simulations.

The proposed workflow — from CFD simulation through statistical analysis to data-driven symbolic model discovery — offers a reproducible and generalizable framework applicable to other large-scale mixing systems. Future work will extend the DOE coverage to include additional impeller types and rheological ranges, and validate the surrogate model against pilot-scale experimental data.

5. Acknowledgments

The authors gratefully acknowledge Petrobras (Petróleo Brasileiro S.A.) for their support through Cooperation Agreement No. 5850.0106053.17.9. The authors also thank ANP (National Agency of Petroleum, Natural Gas, and Biofuels) for their support through the regulation of Research and Development (R&D) fees.

6. References

Cranmer, M. (2024). PySR: High-performance symbolic regression in Python and Julia. Astrophysics Source Code Library, ascl:2409.

Fonseca, R. P., Fontoura, D. V. R., Spogis, N., Fonseca, W. D. P., Noriler, D., Castilho, G. J., & Nunhez, J. R. (2025). Development of a machine learning-based symbolic regression model for mixing time in large petroleum storage tanks. Chemical Engineering Science, 316, 121903.

Menter, F. R. (1994). Two-equation eddy-viscosity turbulence models for engineering applications. AIAA Journal, 32(8), 1598–1605.

Patankar, S. V. (1980). Numerical heat transfer and fluid flow. Hemisphere Publishing Corporation.

Roache, P. J. (1994). Perspective: A method for uniform reporting of grid refinement studies. Journal of Fluids Engineering, 116(3), 405–413.