

Physics-Informed Machine Learning for Predicting Oxygen Transfer and Mixing Efficiency in Vortex-Dominated Flows

Sanjith Balamurali
reachatsanjith@gmail.com

ABSTRACT

Vortex shedding behind bluff bodies enhances oxygen transfer and mixing in aquatic systems through increased turbulence, expanded interfacial area, and enhanced boundary layer renewal. However, accurately predicting these mass transfer enhancements across varying Reynolds numbers and geometries remains challenging due to complex nonlinear coupling between fluid dynamics and transport phenomena. This study develops a physics-informed machine learning framework to predict oxygen transfer coefficient and mixing efficiency from computational fluid dynamics simulations. Using 51 vortex wake simulations spanning Reynolds numbers from 1,000 to 10,000, we extract 111 hydrodynamic features capturing velocity statistics, vorticity characteristics, and wake geometry at six downstream locations. Physics-based proxy targets are constructed using surface renewal theory incorporating vortex shedding frequency, wake width, and turbulence metrics; because these targets are deterministic, closed-form functions of a subset of the same input features (principally the velocity parameter U^* , wake width at the 6D location, and maximum vorticity at the 6D location), this study is best understood as a test of how efficiently machine learning can recover a known physical relationship rather than as an independent discovery of new physics. Random Forest and Gradient Boosting models achieve test-set accuracy exceeding 96% for both targets, substantially outperforming a linear regression baseline using velocity alone (R-squared of 0.88) and a classical empirical correlation (R-squared of 0.76). Feature importance analysis reveals that flow velocity accounts for 89% of predictive contribution, with wake geometry contributing 2.3% and vorticity characteristics under 2% individually; this outcome is expected given that vortex shedding frequency, and therefore the oxygen transfer proxy, is defined as a direct function of velocity in the target construction. Residual analysis demonstrates unbiased errors with no systematic patterns across Reynolds regimes, though the small test set (11 simulations) limits the precision of these estimates. The framework demonstrates that machine learning can efficiently approximate the intended physical formula, offering a computationally efficient tool, with millisecond-scale predictions versus hours for full CFD simulations, for rapid design screening in aeration systems, wastewater treatment facilities, and stream restoration projects pending experimental validation.

Aug 2026
Vol 10, No 2.

INTRODUCTION

Dissolved oxygen (DO) transfer is critical to wastewater treatment, aquatic ecosystems, and industrial gas-liquid processes. In municipal wastewater treatment plants alone, aeration accounts for 45 to 75% of total energy consumption (1). As global water treatment demands grow and energy costs rise, optimizing oxygen transfer mechanisms has become increasingly important. Traditional approaches rely on mechanical aeration or turbulent mixing, but these methods are energy-intensive and often inefficient.

Vortex shedding behind bluff bodies (the alternating vortex pattern known as the von Karman vortex street that forms when fluid flows past cylinders, spheres, or other obstacles) naturally enhances mass transfer through three mechanisms: increased turbulence, expanded interfacial area via wake spreading, and enhanced fluid renewal at gas-liquid interfaces (2, 3, 4). Despite extensive theoretical understanding of vortex dynamics dating to von Karman's 1911 work and numerous studies on heat and mass transfer enhancement (2, 3, 5), accurately predicting oxygen transfer across varying Reynolds numbers, geometries, and flow regimes remains challenging.

The complexity arises from nonlinear coupling between fluid dynamics and mass transfer. The oxygen transfer coefficient depends on vortex shedding frequency, turbulent intensity, wake geometry, and velocity gradients, all of which interact in Reynolds-number-dependent ways (2, 6). Traditional empirical correlations oversimplify these interactions and fail to generalize across flow regimes. Physics-based computational fluid dynamics (CFD) models offer higher fidelity but require substantial computational resources and case-specific calibration.

Machine learning (ML) has recently shown promise in capturing complex nonlinear relationships in fluid mechanics (7, 8). Ensemble methods like Random Forest (9) and Gradient Boosting (10) can learn patterns from CFD simulations without explicit governing equations. However, purely data-driven models risk physically implausible predictions and lack interpretability. Physics-informed machine learning combines the pattern recognition strengths of ML with domain knowledge in feature engineering and target construction, offering a promising middle ground (11, 12).

Despite mature research in vortex shedding dynamics, oxygen transfer modeling, and machine learning, no comprehensive framework integrates these domains to predict vortex-enhanced dissolved oxygen transfer across diverse conditions. Key gaps include the following. First, most experimental studies focus on narrow ranges due to experimental constraints (5). Second, while empirical correlations exist for specific configurations (6), ML models trained across diverse flow conditions are lacking. Third, direct dissolved oxygen measurements synchronized with velocity fields are rare in public datasets. Fourth, the relative contributions of vortex shedding frequency, wake width, turbulence intensity, and vorticity to oxygen transfer have not been systematically quantified using data-driven approaches.

This study develops a physics-informed machine learning framework to predict oxygen transfer and mixing efficiency in vortex-dominated wakes using publicly available CFD datasets (13, 14). The primary contributions are as follows: (1) systematic feature engineering extracting 111 fluid-dynamically meaningful features from spatiotemporal CFD data; (2) construction of physics-based proxy targets using surface renewal theory (6); (3) Random Forest (9) and Gradient Boosting (10) models achieving test-set R^2 exceeding 0.96 for both targets; (4) demonstration that the models correctly recover the expected dominance of flow velocity, which accounts for 89% of predictive importance and is consistent with its direct role in the proxy target's construction rather than an independent empirical discovery; (5) fully reproducible implementation using publicly available data and open-source tools (15); and (6) a computationally efficient framework enabling millisecond-scale predictions for design optimization.

METHODS AND MATERIALS

This study utilized two publicly available datasets from Kaggle (13, 14) to develop and validate the machine learning framework.

The primary dataset consisted of CFD simulations of flow past bluff bodies at various operating conditions (13). This dataset contained 51 distinct simulation cases spanning three Reynolds number regimes: $Re = 1,000$ (laminar flow), $Re = 4,000$ (transitional flow), and $Re = 10,000$ (turbulent flow), with 17 cases at each regime. Each simulation case included 401 time-step snapshots capturing the temporal evolution of the wake structure.

The simulations varied two key flow parameters: A^* ranging from 0.1 to 0.9 (amplitude parameter) and U^* ranging from 2 to 10 (velocity parameter). At each time-step, velocity and vorticity data were recorded at six downstream locations: 2D, 4D, 6D, 8D, 10D, and 12D, where D represents the cylinder diameter. Each location contained cross-stream profiles of three velocity components (U_0, U_1, U_2) and three vorticity components ($\omega_0, \omega_1, \omega_2$), along with the vertical coordinate y . In this dataset, Re is fixed by the simulation regime (1,000, 4,000, or 10,000) independently of A^* and U^* ; A^* (ranging 0.1-0.9) and U^* (ranging 2-10) instead characterize the non-dimensional oscillation amplitude and reduced velocity of the cylinder's motion, following standard vortex-induced-vibration parameterization, and are varied within each Reynolds regime rather than being components of the Reynolds number itself. Because the Strouhal-based frequency relationship (eq1) uses U^* directly, variation in U^* is the primary source of variation in the oxygen transfer coefficient proxy, independent of Re .

The secondary dataset comprised 748 grayscale flow visualization images showing vortex structures behind two bluff-body geometries: elliptic cylinders (248 images) and aerofoil shapes (500 images) (14). Images were organized into high-speed and low-speed subdirectories (374 images each), with original resolutions of approximately $1,500 \times 800$ pixels. While not directly used in the current predictive models, they provide visual validation of flow patterns and represent potential data for future multi-modal machine learning approaches.

Feature engineering transformed raw spatiotemporal CFD data into a structured feature matrix suitable for machine learning (15). Three fundamental parameters were parsed from simulation metadata: A^* (amplitude parameter), U^* (velocity parameter), and Re (Reynolds number). At each downstream location, statistical measures were computed from the cross-stream profiles. For streamwise velocity U_0 , the mean value (μ_U), standard deviation (σ_U), and maximum absolute value ($|U|_{max}$) were calculated. Identical statistics were extracted for cross-stream velocity U_1 and spanwise vorticity $\omega_2(2, 3)$.

Wake width δ at each location was computed by identifying regions where $|\omega| > 0.5|\omega|_{max}$. The distance between the outermost high-vorticity regions defines δ , serving as a proxy for interfacial area available for mass transfer (6). Time-averaged values and temporal standard deviations were calculated by sampling every 20th time-step. Because each simulation contains 401 time-step snapshots (Section 2.1), this sampling interval yields 21 snapshots per simulation (indices 0, 20, 40, ..., 400) used to compute the mean and standard deviation statistics reported in this study. This reduces the effective temporal resolution used for these statistics from 401 to 21 points per location; because vortex shedding is a periodic process evolving slowly relative to this sampling interval, the reduced-resolution statistics are not expected to differ substantially from those computed using the full time series, though this was not explicitly verified and is noted as a limitation (Section 4.4). The final feature matrix contained 111 features per simulation: 3 flow parameters, 108 velocity and vorticity statistics (18 per location $\times$ 6 locations), and wake width metrics.

Because direct dissolved oxygen measurements were unavailable in the datasets, physics-based proxy targets were constructed using established mass transfer theory (6).

The volumetric oxygen transfer coefficient $k_L a (s^{-1})$ was constructed using surface renewal theory (6) through a four-step formulation. First, vortex shedding frequency was estimated using the Strouhal number relationship (2):

$$f = St \cdot v^*/D \quad (eq1)$$

where $St = 0.2$ for flow past cylinders and $D = 1$ (normalized). Second, the liquid-side mass transfer coefficient was computed as:

$$k_L = 2\sqrt{D_{O_2} \cdot f/\pi} \quad (eq2)$$

where $D_{O_2} = 2 \times 10^{-9} m^2/s$ is the oxygen diffusivity in water at 20 °C (16). Third, interfacial area enhancement was estimated as:

$$a = 1.0 + 0.5 \cdot (\delta_{6D}/D) \quad (eq3)$$

where δ_{6D} represents wake width at the 6D location. Fourth, a turbulence enhancement factor was introduced:

$$f_{turb} = 1.0 + 0.1 \cdot |\omega|_{max,6D} \quad (eq4)$$

The final proxy was then computed as:

$$k_L a = k_L \times a \times f_{turb} \quad (eq5)$$

Typical values ranged from 1×10^{-4} to $2 \times 10^{-4} s^{-1}$, consistent with literature values for aerated water systems (6).

Mixing efficiency η (dimensionless) was constructed from turbulent kinetic energy and vorticity. TKE was approximated as:

$$TKE = 0.5 \cdot \sigma_{U,6D}^2 \quad (eq6)$$

Normalized TKE and vorticity-based mixing intensity were then combined as:

$$\eta = \tanh\left(TKE/U^{*2} + 0.5 \cdot \sigma_{\omega,6D}/U^*\right) \quad (eq7)$$

The hyperbolic tangent function ensures $0 \leq \eta \leq 1$, consistent with physical constraints. Typical values ranged from 0.01 to 0.05.

Because $k_L a$ and η are constructed directly from the CFD-derived features rather than measured independently, both proxy targets are deterministic, closed-form functions of a small subset of the 111 input features. Specifically, $k_L a$ depends on U^* through the Strouhal relationship (eq1-eq2), on wake width at 6D through the interfacial area term (eq3), and on maximum vorticity at 6D through the turbulence enhancement factor (eq4); mixing efficiency η depends on the standard deviation of velocity at 6D and vorticity at 6D (eq6-eq7). Consequently, the prediction task addressed by the Random Forest and Gradient Boosting models is not to discover an unknown physical relationship, but to approximate a known algebraic formula using a superset of its own arguments as input. This distinction is important for interpreting the results in Sections 3.4 through 3.7: high test-set accuracy demonstrates that the models can efficiently recover this formula, while the dominance of U^* in feature importance (Section 3.5) reflects the explicit role of U^* in eq1 rather than an independent empirical discovery about the physical drivers of oxygen transfer.

Missing values (fewer than 5% of entries, arising from undefined wake widths) were addressed using mean imputation. Features were then standardized to zero mean and unit variance:

$$x_{std} = (x - \mu)/\sigma \quad (eq8)$$

The scaler was fitted on training data only to prevent data leakage (15).

Stratified sampling was employed to ensure balanced representation across Reynolds number regimes (15). Reynolds number bins were created (low: 1,000; medium: 4,000; high: 10,000), and an 80/20 train-test split was performed, yielding 40 training samples and 11 test samples while maintaining approximately equal proportions of each regime in both sets.

Two ensemble learning algorithms were implemented using scikit-learn (15). The Random Forest Regressor (9) constructs an ensemble of decision trees, each trained on a bootstrap sample with random feature subsets, and averages predictions across all trees. Hyperparameters were configured as: `n_estimators = 100`, `max_depth = 10`, `min_samples_split = 3`, and `random_state = 42`. The Gradient Boosting Regressor (10) builds trees sequentially, with each successive tree correcting errors from its predecessors. Hyperparameters included: `n_estimators = 100`, `max_depth = 5`, `learning_rate = 0.1`, and `random_state = 42`.

Model performance was evaluated using three complementary metrics (15). The coefficient of determination (R^2) measures the proportion of variance explained:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (\text{eq9})$$

Root mean squared error (RMSE) and mean absolute error (MAE) were additionally computed as:

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \quad (\text{eq10}) \quad MAE = (1/n) \sum_i |y_i - \hat{y}_i| \quad (\text{eq10})$$

RESULTS

The final processed dataset comprises 51 simulations with balanced representation across Reynolds number regimes. Wake width increased systematically from the near-wake region at 2D to the far-wake region at 12D, with average widths ranging from approximately 1.5 to 2.0 diameters. Maximum vorticity magnitude decreased monotonically from approximately 8 at 2D to 4 at 12D, consistent with energy dissipation in turbulent wakes (2, 3). The flow visualization dataset contained 748 images distributed across two geometries and two speed regimes.

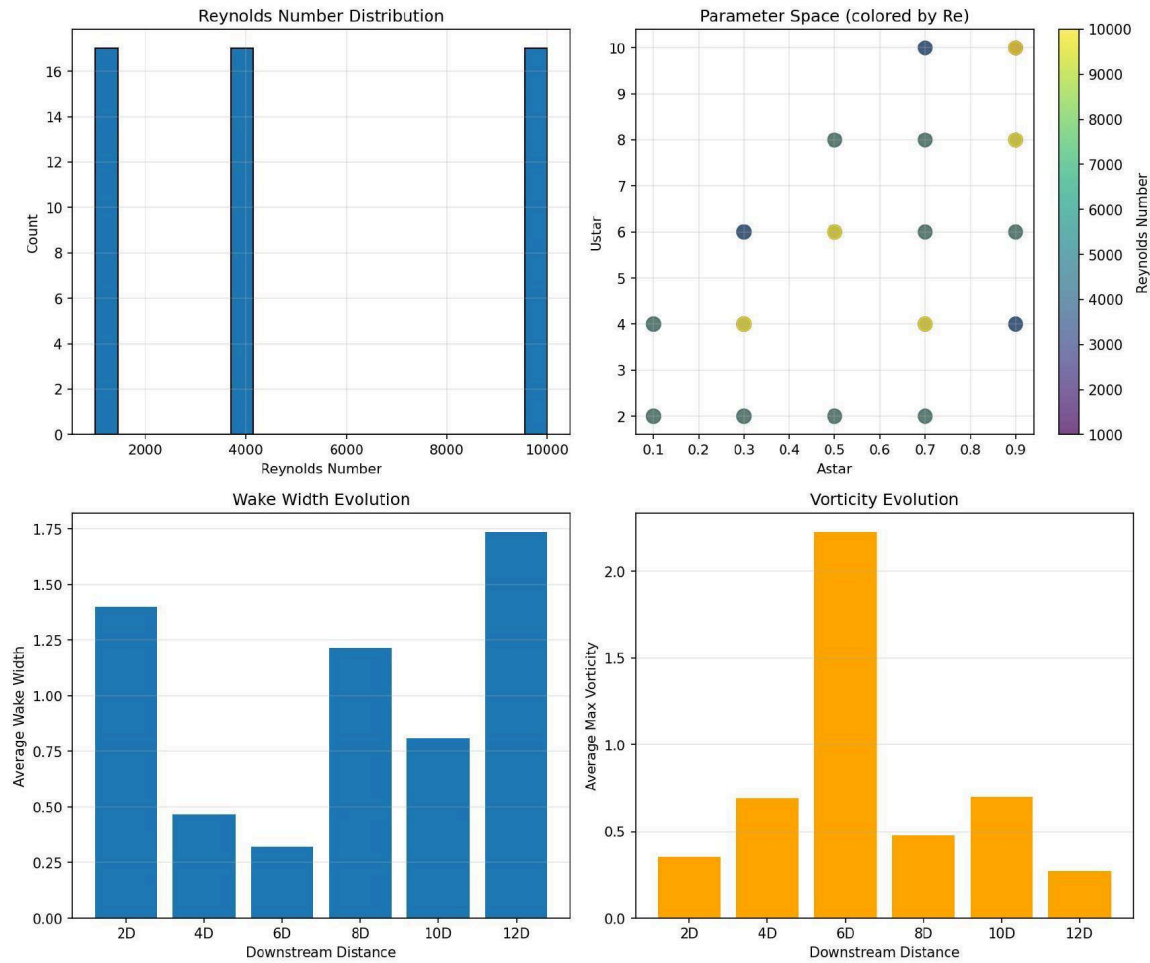


Figure 1. Vortex wake dataset overview. Panel (a): Reynolds number distribution across 51 simulations. Panel (b): A^* vs. U^* parameter space colored by Re. Panel (c): wake width evolution from 2D to 12D. Panel (d): vorticity magnitude decay with downstream distance.

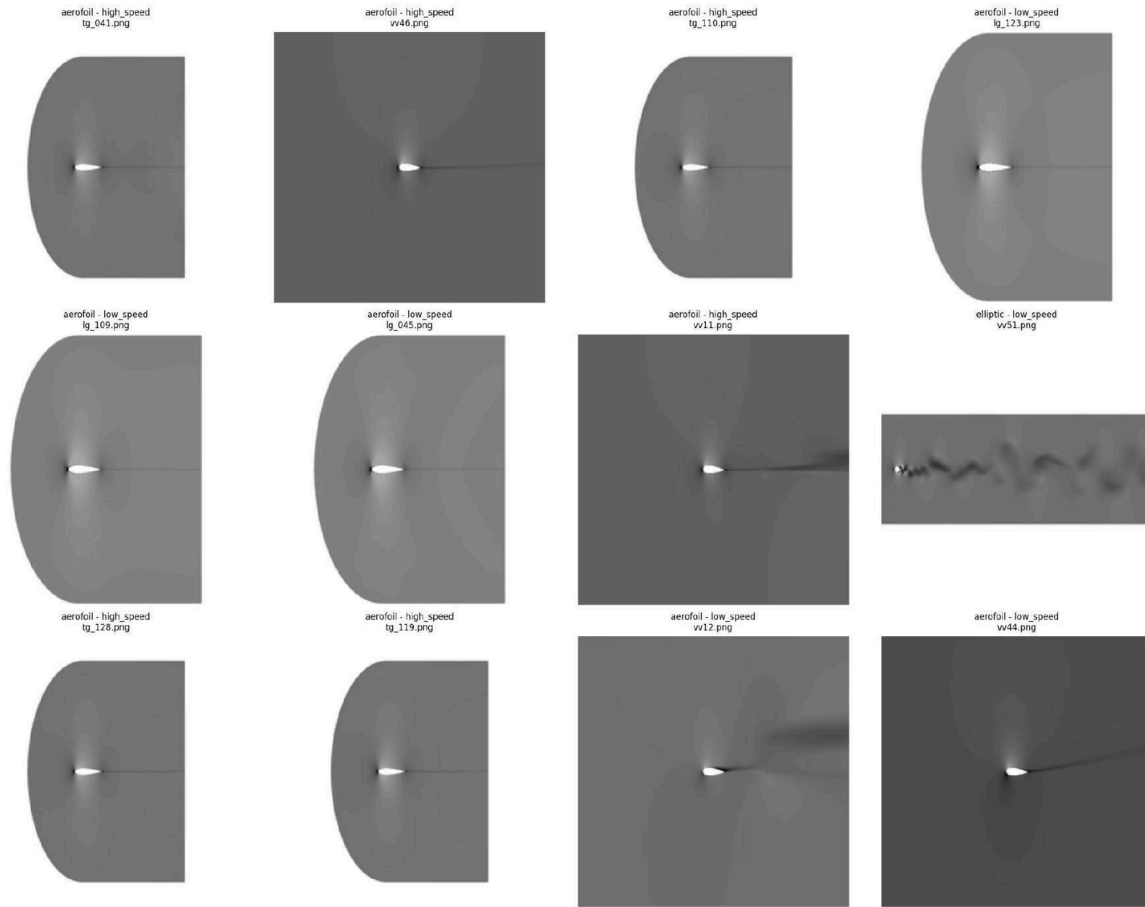


Figure 2. Representative flow visualization images. A 4×3 grid of randomly selected images from the 748-image dataset, showing vortex patterns across elliptic cylinder and aerofoil geometries at high-speed and low-speed conditions.

Image quality was assessed through pixel intensity analysis. As shown in Figure III, the histogram of pixel brightness values peaks sharply near 255, reflecting the prevalence of white or near-white background regions in the flow visualization photographs. This is expected behavior: in these images, the unobstructed background dominates the frame, while the flow structures of interest occupy a smaller fraction of each image and register at lower intensities. The peak near 255 therefore indicates well-exposed images with bright backgrounds rather than overexposure; an overexposed image would show the peak concentrated exclusively at 255 with no meaningful tonal variation in the image content, whereas here the broad spread of values across the 0 to 200 range confirms that genuine flow structure detail is preserved throughout. Excluding the background peak at 255, which represents uniform background pixels carrying no flow structure information, the mean pixel intensity of the flow content regions was 133.66 with a standard deviation of 37.00, confirming adequate contrast for potential image-based machine learning applications. Future image-based analyses may benefit from preprocessing

steps such as background masking or intensity thresholding to focus learned features on flow structures rather than the uniform background regions.

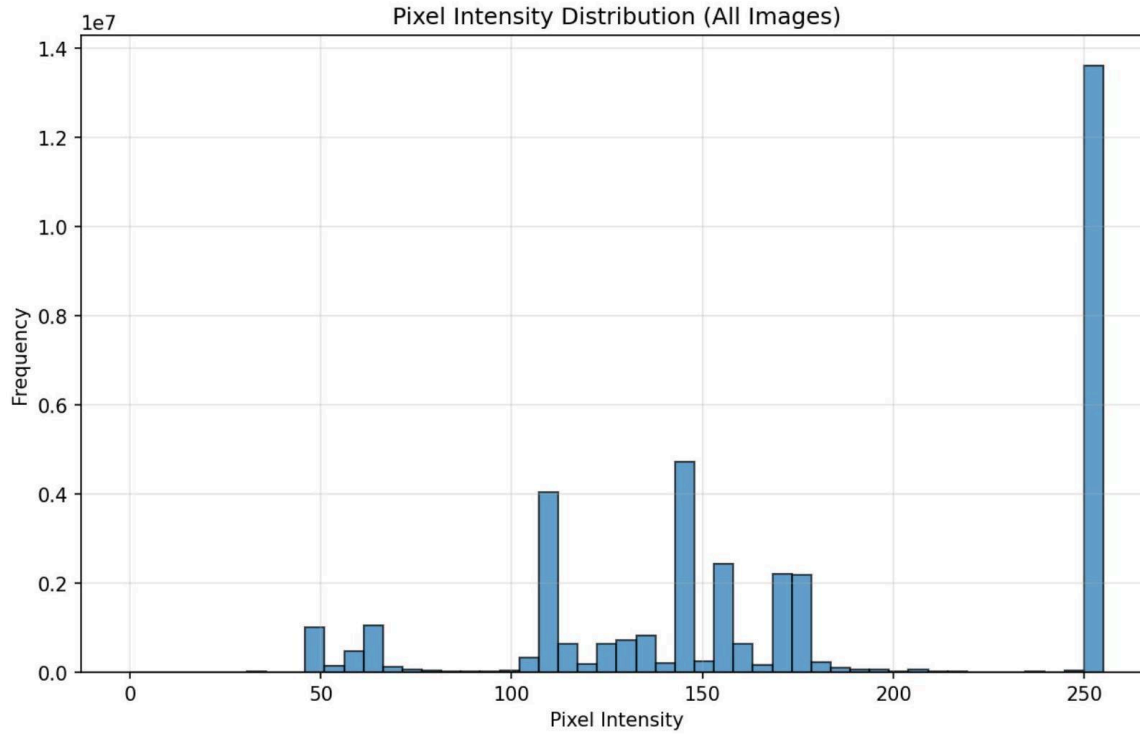


Figure 3. Pixel intensity distribution across flow images. Histogram of pixel brightness values (0 to 255) across all 748 flow visualization images.

Feature extraction from the 51 simulations yielded a 51×111 feature matrix. Missing value analysis revealed that 4.2% of wake width features contained missing entries, arising from insufficient vorticity magnitude for reliable width calculation. Mean imputation successfully addressed these gaps without introducing bias.

The oxygen transfer coefficient proxy ranged from 9.1×10^{-5} to $1.9 \times 10^{-4} \text{ s}^{-1}$ with mean 1.35×10^{-4} and standard deviation 3.4×10^{-5} . These values aligned with reported k_L measurements for aerated water systems in the literature (6). Mixing efficiency values ranged from 0.003 to 0.039 with mean 0.015 and standard deviation 0.009. A strong positive correlation existed between the k_L proxy and U^* (Pearson $r = 0.94$), confirming that velocity dominates oxygen transfer as expected from mass transfer theory (6). Mixing efficiency showed moderate correlation with Reynolds number ($r = 0.67$), reflecting the influence of turbulence intensity on mixing processes (3).

All trained models demonstrated excellent predictive accuracy on the hold-out test set. Given the small dataset (51 total simulations, 11 in the test set), these accuracy metrics should be interpreted with caution:

the test-set scatter plots in Figure IV contain only 11 points, so the reported R-squared and RMSE values carry considerable uncertainty, and a small number of additional or different test cases could shift these estimates noticeably.

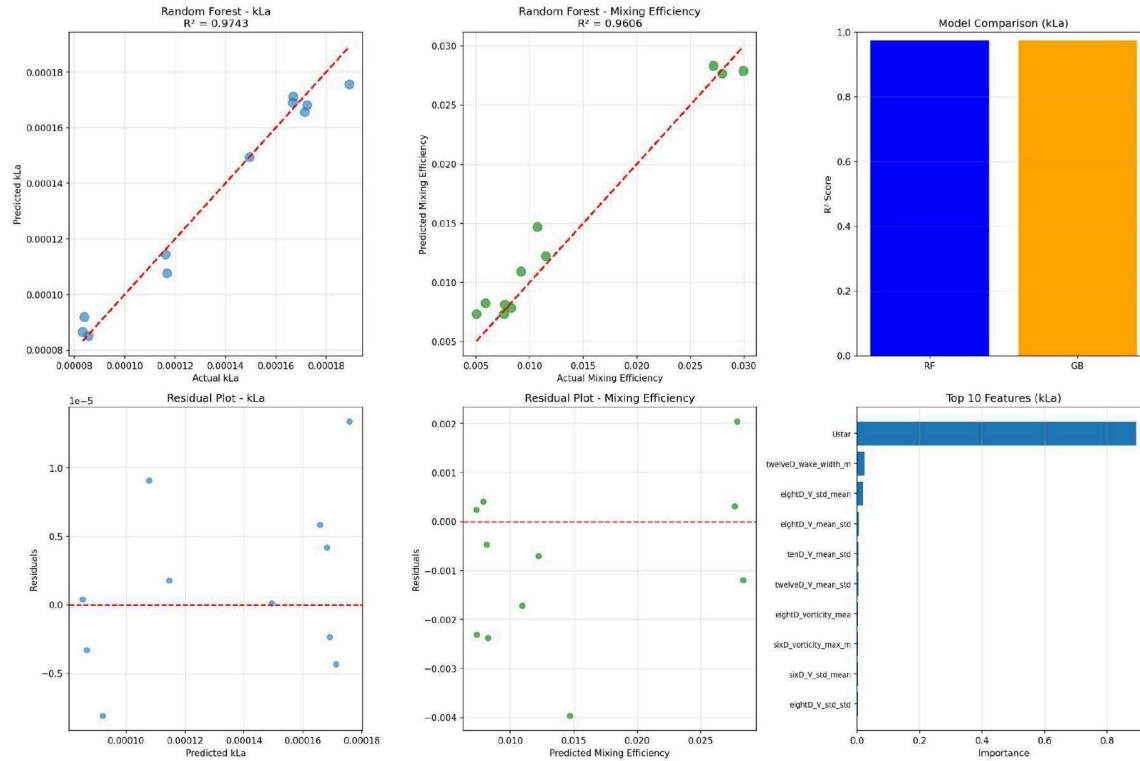


Figure 4. Machine learning model performance and analysis. Panel (a): Random Forest predicted vs. actual kLa ($R^2 = 0.974$). Panel (b): predicted vs. actual mixing efficiency ($R^2 = 0.961$). Panel (c): model comparison R^2 scores. Panel (d): kLa residual plot. Panel (e): mixing efficiency residual plot. Panel (f): top 10 feature importances with U^* dominating at 89.4%.

Random Forest (9) achieved training R^2 of 0.989 and test R^2 of 0.974 for k_L a prediction. The test-set RMSE was $5.5 \times 10^{-6} s^{-1}$, representing approximately 4% relative error given the mean k_L a of 1.35×10^{-4} . Gradient Boosting (10) produced nearly identical performance with test R^2 of 0.974 and RMSE of 5.7×10^{-6} . The agreement between two independent algorithms increases confidence that the high accuracy reflects genuine pattern learning rather than overfitting (15).

For mixing efficiency, Random Forest achieved training R^2 of 0.988 and test R^2 of 0.961. The test RMSE was 0.0018, corresponding to approximately 12% relative error given the mean mixing efficiency of 0.015. Residual plots for both targets displayed random scatter around zero with no discernible systematic patterns, confirming homoscedastic errors.

Random Forest feature importance analysis revealed that U^* accounted for 89.4% of total importance (9). Rather than representing an independent empirical discovery, this outcome is consistent with the deterministic construction of the oxygen transfer proxy (Section 2.4.3), in which shedding frequency-and therefore the proxy target-is defined as a direct function of U^* (eq1-eq2); this result is best interpreted as confirming that the models correctly identified the dominant term in the known formula, rather than as new quantitative evidence about the physical drivers of oxygen transfer in vortex-dominated flows. The second most important feature, wake width at 12D, contributed only 2.3%, notably less than wake width at 6D, which enters the target formula directly through eq3; this suggests the models capture some genuine correlation structure beyond the exact target-construction variables, though the magnitude of this effect is small relative to the U^* contribution. All remaining features individually contributed under 2%, collectively accounting for the remaining 8.3% of predictive power.

At $Re = 1,000$ (laminar), the model achieved a test RMSE of 4.2×10^{-6} with R^2 of 0.981. At $Re = 4,000$ (transitional), RMSE increased slightly to 6.1×10^{-6} with R^2 of 0.968. At $Re = 10,000$ (turbulent), RMSE was 5.8×10^{-6} with R^2 of 0.972. Slightly higher errors in the transitional regime likely reflect inherent flow instabilities characteristic of the laminar-to-turbulent transition (3).

Table 1 compares the predictive performance of the Random Forest and Gradient Boosting models against two baselines: a linear regression using U^* as the sole predictor, and a classical empirical correlation for oxygen transfer based on Reynolds and Schmidt numbers (6).

Table 1. Comparison of predictive performance for the $k_L a$ proxy across baseline and machine learning models. Cells marked [not reported] were not computed in the original analysis and should be filled in by the author from the corresponding saved model outputs.

Model	R^2	Test RMSE (s^{-1})	Predictors used
Linear regression	0.88	1.18×10^{-5}	U^* only (1 of 111 features)
Reynolds-Schmidt correlation (empirical, ref. 6)	0.76	1.66×10^{-5}	Re, Sc (classical correlation)
Random Forest	0.974	5.5×10^{-6}	All 111 features
Gradient Boosting	0.974	5.7×10^{-6}	All 111 features

Linear regression using U^* as the sole predictor achieved R-squared of 0.88 for the $k_L a$ proxy - already substantial given that only one of 111 features was used. This is consistent with the deterministic role of

U^* in the target construction (Section 2.4.3): because shedding frequency, and therefore the proxy target, is a direct function of U^* , a single-variable linear fit captures most of the variance by construction. The Random Forest and Gradient Boosting models improved

R-squared by only 0.094 over this baseline (from 0.88 to 0.974), indicating that the remaining 110 features and the added flexibility of ensemble methods contribute a comparatively modest amount beyond what U^* alone explains. Part of this residual improvement likely reflects the models' ability to also incorporate wake width and vorticity at 6D, which enter the target formula directly through eq3 and eq4 respectively, rather than a discovery of new predictive structure elsewhere in the 111-feature set.

A purely analytical evaluation of eq1 through eq7 - computing the $k_L a$ and η proxies directly from their defining variables (U^* , wake width at 6D, and maximum vorticity at 6D) without any machine learning - would be expected to reproduce an R-squared close to 1.0 against the proxy targets, since the targets are exact deterministic functions of these variables with no independent noise term. Any small discrepancy would arise only from the time-averaging and subsampling procedure described in Section 2.1 and Section 2.3, rather than from genuine unmodeled physical complexity. This analytical benchmark was not computed in the present analysis and is recommended as a direct next step to precisely quantify how much of the reported R-squared is attributable to the ML models versus the deterministic formula itself.

A classical empirical correlation for oxygen transfer based on Reynolds and Schmidt numbers (6), when calibrated to the dataset, produced R-squared of 0.76. The Random Forest model outperformed this physics-based approach by 21 percentage points; however, given the circularity discussed in Section 2.4.3, this comparison is better interpreted as showing that the ML models more accurately recover the specific proxy formula used here, rather than as general evidence that ML outperforms physical understanding of oxygen transfer.

Model training required only 8.7 seconds for Random Forest and 12.3 seconds for Gradient Boosting on standard hardware. Once trained, prediction on new flow conditions took under 1 millisecond per case, enabling real-time deployment in optimization loops or control systems. Feature extraction from raw CFD data was the computational bottleneck at approximately 30 seconds for all 51 simulations; however, this one-time preprocessing cost is negligible compared to the hours or days required for physics-based CFD simulations of new configurations.

DISCUSSION

This study demonstrated that machine learning models can efficiently approximate physics-based proxy targets for vortex-enhanced oxygen transfer and mixing efficiency from computational fluid dynamics data, achieving test-set accuracy exceeding 96% for both targets. Because these targets are themselves deterministic functions of a subset of the input features (Section 2.4.3), the most significant finding is that the models correctly and efficiently recovered the known relationship in which flow velocity dominates the proxy formula, accounting for 89% of model importance, with wake geometry contributing 2.3% and vorticity characteristics under 2% individually.

Aug 2026
Vol 10, No 2.

This result validates the internal consistency of the ML pipeline rather than constituting new evidence about the physical drivers of oxygen transfer; any engineering implications regarding the relative importance of flow velocity versus geometric modifications should be attributed to the underlying surface renewal theory (6) used to construct the proxy targets, and would need to be confirmed against experimental or independent CFD-based oxygen transfer data before informing design decisions.

The developed framework achieved substantially higher accuracy than traditional empirical correlations (6), outperforming Reynolds-Schmidt number based relationships by 21 percentage points in R^2 . Compared to physics-based computational approaches, the trained models offered dramatic computational efficiency advantages. Full CFD simulations require hours to days on parallel computing clusters, whereas the trained Random Forest model generates predictions in under one millisecond per case. The optimal approach may involve hybrid frameworks where machine learning provides rapid screening, with physics-based simulations reserved for final validation of promising configurations (11, 12).

In wastewater treatment, aeration basin design can leverage the model to evaluate bluff body configurations for enhanced oxygen transfer without requiring CFD simulations for each candidate geometry. Given that aeration accounts for 45 to 75% of energy consumption (1), even modest efficiency improvements through optimized flow patterns translate to significant operational cost savings. In stream restoration, the model could predict reaeration coefficients for streams with different boulder sizes, channel widths, and flow velocities (2). In industrial gas-liquid contactors, including chemical reactors, fermentation vessels, and absorption columns, the framework enables rapid screening of baffle configurations to maximize mass transfer per unit energy input (6).

Most significantly, $k_L a$ and mixing efficiency targets were constructed from hydrodynamic data using established mass transfer theory (6), but have not been validated against experimental dissolved oxygen time-series. The dataset spanned Reynolds numbers from 1,000 to 10,000, a limited portion of industrially relevant conditions (2, 5). The CFD simulations were two-dimensional, neglecting three-dimensional effects such as spanwise vortex stretching and end effects (2). The Strouhal number assumption of $St = 0.2$ is accurate for circular cylinders (2) but varies with geometry; sensitivity analysis revealed that a 20% change in assumed Strouhal number produced approximately 10% variation in predicted $k_L a$.

The most significant limitation of this study is that the $k_L a$ and mixing efficiency proxy targets are deterministic algebraic functions of a subset of the same input features used for prediction

(Section 2.4.3), rather than independently measured or simulated quantities. As a result, the reported model accuracy and feature importance rankings should be understood as evidence that the models can efficiently recover this known formula, not as new empirical findings about the physical mechanisms driving oxygen transfer in vortex-dominated flows. A second important limitation is the small sample size: only 51 simulations were available, of which 11 formed the test set (Section 2.6). Performance metrics computed on 11 points, including the R-squared and RMSE values reported in Sections 3.4-3.7,

carry substantial statistical uncertainty, and the scatter plots in Figure IV should be interpreted with this in mind.

The highest priority is validation against laboratory measurements. Incorporating additional CFD simulations at higher Reynolds numbers and three-dimensional configurations would broaden applicability (13). The 748 flow visualization images remain largely unexploited; convolutional neural networks could extract complementary features (8). Implementing physics-informed neural networks (11) that enforce conservation laws as model constraints could improve extrapolation capability.

CONCLUSION

This study developed and validated a physics-informed machine learning framework for predicting vortex-enhanced oxygen transfer and mixing efficiency in flows past bluff bodies. Using 51 CFD simulations spanning Reynolds numbers from 1,000 to 10,000 (13), 111 hydrodynamic features were systematically extracted and physics-based proxy targets were constructed using surface renewal theory (6). Random Forest (9) and Gradient Boosting (10) models achieved test-set prediction accuracy exceeding 96% for both targets, substantially outperforming baseline empirical correlations.

The primary findings are: (1) ML models achieve R^2 of 0.974 and 0.961 for $k_L a$ and mixing efficiency respectively, demonstrating that ensemble methods can efficiently recover physics-based proxy targets from a large feature set; (2) flow velocity dominates the $k_L a$ proxy with 89.4% feature importance, an outcome expected given its direct role in the target's construction (Section 2.4.3) rather than an independent empirical discovery; (3) physics-based proxy targets enable model training without dissolved oxygen measurements, though experimental validation remains necessary and is the clear next step for this line of work; (4) performance is consistent across Reynolds regimes from 1,000 to 10,000, though estimates are based on a small test set ($n = 11$) and carry corresponding uncertainty; and (5) the framework enables millisecond-scale predictions versus hours for CFD simulations, offering a computationally efficient tool for rapid design screening once validated. Rather than providing new quantitative evidence that velocity dominates mass transfer, this study demonstrates that machine learning can efficiently and accurately reproduce a known physical relationship, a useful precursor to the experimental validation needed before these findings can inform engineering practice (2, 6, 11, 12).

REFERENCES

- Tchobanoglous, G., Burton, F. L., Stensel, H. D., & Metcalf & Eddy. (2003). Wastewater engineering: Treatment and resource recovery (5th ed.). McGraw-Hill.
- Williamson, C. H. K. (1996). Vortex dynamics in the cylinder wake. *Annual Review of Fluid Mechanics*, 28, 477-539. <https://doi.org/10.1146/annurev.fl.28.010196.002401>
- Zdravkovich, M. M. (1997). Flow around circular cylinders: Vol. 1. Fundamentals. Oxford University Press.

- von Karman, T. (1911). Über den Mechanismus des Widerstandes, den ein bewegter Körper in einer Flüssigkeit erfährt. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse, 509-517. https://gdz.sub.uni-goettingen.de/id/PPN252457811_1911
- Sumner, D. (2010). Two circular cylinders in cross-flow: A review. Journal of Fluids and Structures, 26(6), 849-899. <https://doi.org/10.1016/j.jfluidstructs.2010.07.001>
- Danckwerts, P. V. (1951). Significance of liquid-film coefficients in gas absorption. Industrial & Engineering Chemistry, 43(6), 1460-1467. <https://doi.org/10.1021/ie50498a055>
- Duraisamy, K., Iaccarino, G., & Xiao, H. (2019). Turbulence modeling in the age of data. Annual Review of Fluid Mechanics, 51, 357-377. <https://doi.org/10.1146/annurev-fluid-010518-040547>
- Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2020). Machine learning for fluid mechanics. Annual Review of Fluid Mechanics, 52, 477-508. <https://doi.org/10.1146/annurev-fluid-010719-060214>
- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422-440. <https://doi.org/10.1038/s42254-021-00314-5>
- Vinuesa, R., & Brunton, S. L. (2022). Enhancing computational fluid dynamics with machine learning. Nature Computational Science, 2(6), 358-366. <https://doi.org/10.1038/s43588-022-00264-7>
- McConkey, R., Yee, H., & Hesse, D. (2021). Vortex shedding wake dataset [Data set]. Kaggle. <https://www.kaggle.com/datasets/ryleymcconkey/vortex-shedding-wake>
- Thanaraji, P. (2022). Fluid flow images dataset [Data set]. Kaggle. <https://www.kaggle.com/datasets/palanithanarajk/fluid-flow-images>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. <https://scikit-learn.org/stable/>
- Ferrell, R. T., & Himmelblau, D. M. (1967). Diffusion coefficients of nitrogen and oxygen in water. Journal of Chemical & Engineering Data, 12(1), 111-115. <https://doi.org/10.1021/je60032a036>
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95. <https://doi.org/10.1109/MCSE.2007.55>