

PROFILE-AWARE PREDICTION AND RESOURCE-CONSTRAINED OPTIMIZATION OF MALACHITE GREEN PHOTODEGRADATION USING Pd–BiFeO₃

Vladan Nedelkovski^{1a}, Sonja Stanković^{1b}, Dragana Medić^{1c},
Aleksandar Cvetković^{1d}, Milan Radovanović^{1e}, Srećko Stopić², Žaklina Tasić^{1f}

¹Technical Faculty in Bor, University of Belgrade, V.J. 12, 19210 Bor, Serbia

²IME Process Metallurgy and Metal Recycling, RWTH Aachen University, Intzestrassse 3, 52056 Aachen, Germany

^{1a} vnedelkovski@tfbor.bg.ac.rs, <https://orcid.org/0000-0001-7724-1118>,

^{1b} sstankovic@tfbor.bg.ac.rs, <https://orcid.org/0000-0003-0745-5422>,

^{1c} dmedic@tfbor.bg.ac.rs, <https://orcid.org/0000-0001-9980-5949>,

^{1d} acvetkovic@tfbor.bg.ac.rs, <https://orcid.org/0009-0003-5184-6874>,

^{1e} mradovanovic@tfbor.bg.ac.rs, <https://orcid.org/0000-0002-5175-6022>,

^{1f} ztasic@tfbor.bg.ac.rs, <https://orcid.org/0000-0001-6544-1980>,

² sstopic@ime-aachen.de, <https://orcid.org/0000-0002-1752-5378>

Abstract

Machine-learning models for photocatalysis can appear overly accurate when time points from the same experiment enter training and test subsets. A 2 wt% Pd–BiFeO₃/malachite-green dataset was aggregated to 200 profile-time observations and evaluated using row-wise and profile-aware five-fold validation. CatBoost reached $R^2 = 0.983$ on the row-wise hold-out, whereas support-vector regression gave the strongest grouped generalization (mean $R^2 = 0.865$). Discrete constrained screening indicated that increasing catalyst loading from 1.0 to 1.5 g L⁻¹ reduced the predicted time to 90% degradation from 242 to 210 min and nominal irradiation-energy input by 13.4%. Profile-aware validation is essential for credible optimization.

Keywords: machine learning, grouped validation, photocatalysis, malachite green, optimization

1. INTRODUCTION

Heterogeneous photocatalysis with oxide-based nanomaterials is regarded as a promising advanced oxidation route for removing persistent toxic, mutagenic, and carcinogenic pollutants from water [1]. BiFeO₃-based materials are attractive because their visible-light response can be modified through noble-metal incorporation. Malachite green (MG) is particularly relevant as a strongly colored and potentially hazardous triphenylmethane dye. Jaffari et al. experimentally evaluated noble-metal-modified BiFeO₃ materials and constructed a machine-learning databank for MG degradation under varied material and operating conditions [2]. Data-driven models can capture nonlinear relationships among catalyst loading, irradiation intensity, reaction time, pH, and water-matrix composition. Artificial-neural-network approaches have been widely applied to photocatalytic degradation modelling, sensitivity analysis, and operating-parameter optimization [3]. The original study reported a test R^2 of approximately 0.99 for CatBoost using 1200 observations [2,4]. However, kinetic

datasets contain repeated measurements from the same experimental profile. A conventional row-wise split can place neighbouring time points and replicates from one experiment in both training and test subsets, producing optimistic estimates of transfer to genuinely new operating conditions. Leakage-aware separation by experimental units is therefore important in machine-learning-based science [5]. This work compares conventional and profile-aware validation for predicting MG photodegradation over 2 wt% Pd–BiFeO₃. Five regression families were evaluated, and the best profile-generalizing surrogate was used for a resource-constrained search targeting at least 90% degradation with reduced irradiation time and nominal irradiation-energy input. All preprocessing, regression, validation, and screening procedures were implemented using custom-made Python scripts.

2. EXPERIMENTAL

2.1. Dataset and preprocessing

The public dataset accompanying Ref. [2] was obtained from the authors' repository [4]. To isolate operating-condition effects and avoid categorical catalyst confounding, only measurements for 2 wt% Pd–BiFeO₃ without added anions were retained. This selection contained 630 raw records. Replicates measured at identical conditions and irradiation times were averaged, giving 200 observations arranged in 20 independent profiles with ten time points from 0 to 270 min. A profile was defined by the unique combination of catalyst loading, light intensity, pH, humic-acid concentration and initial MG concentration.

Table 1. Regression variables and experimental domain

Variable	Unit	Range
Catalyst loading	g L ⁻¹	0.5–2.5
Light intensity	W	0–105
Irradiation time	min	0–270
Solution pH	–	3.0–9.0
Humic acid (HA)	mg L ⁻¹	0–10
Initial MG concentration	mg L ⁻¹	5–80
Degradation efficiency, Y	%	0–100

2.2. Regression and validation

The six process variables listed in Table 1 were used as inputs (X), and degradation efficiency was used as the response (Y). Final MG concentration was excluded because efficiency is calculated directly from the initial and final concentrations and would therefore cause target leakage. The tested models were quadratic response-surface regression (RSM), radial-basis support-vector regression (SVR) [6], an artificial neural network (ANN), CatBoost (CAT) [7], and XGBoost (XGB) [8]. Continuous inputs were standardized for RSM, SVR, and ANN.

A conventional 20% row-wise hold-out and shuffled five-fold cross-validation reproduced the common comparison workflow. A separate five-fold grouped analysis kept all ten time points from each profile in the same fold and served as a robustness assessment of transfer to unseen profiles. Performance was evaluated using R^2 , root-mean-square error (RMSE), and mean absolute error (MAE). MAPE was not interpreted because each profile contains a zero-efficiency value at $t = 0$.

2.3. Resource-constrained optimization

The profile-aware SVR surrogate was fitted to the aggregated subset and subjected to discrete screening over experimentally supported factor arms. At pH 5.45, $HA = 0 \text{ mg L}^{-1}$, $C_0 = 10 \text{ mg L}^{-1}$, and 105 W, catalyst loading was restricted to observed levels and irradiation time was interpolated at 0.5-min resolution. Candidates were required to reach at least 90% predicted degradation. Following constrained multiobjective optimization principles [9], nominal irradiation-energy input E and catalyst loading L_{cat} were treated as competing objectives; P denotes nominal lamp power, t is irradiation time in minutes, and D_{obs} is the documented domain:

$$\min_{x \in D_{obs}} [E, L_{cat}], \quad E = P \cdot t / 60, \quad \eta_{MG} \geq 0.90 \quad (1)$$

3. RESULTS AND DISCUSSION

3.1. Validation-dependent model performance

Table 2. Conventional and profile-aware predictive performance

Model	Hold-out R^2	Hold-out RMSE	Row CV R^2	Group CV R^2	Group RMSE
RSM	0.9407	7.599	0.940 ± 0.022	0.683 ± 0.207	16.95
SVR	0.9673	5.645	0.971 ± 0.015	0.865 ± 0.084	11.45
ANN	0.9604	6.206	0.947 ± 0.023	0.861 ± 0.092	11.43
XGB	0.9544	6.666	0.958 ± 0.012	0.646 ± 0.162	18.78
CAT	0.9827	4.109	0.976 ± 0.009	0.676 ± 0.177	17.54

All models appeared highly accurate under the conventional row-wise workflow (Table 2). CAT ranked first on the common hold-out ($R^2 = 0.9827$; RMSE = 4.11 percentage points), followed by SVR and ANN. This agrees with the strong CatBoost performance reported for the full material database [2]. The narrow row-wise CV standard deviations further indicate apparently stable interpolation when individual observations are treated as independent.

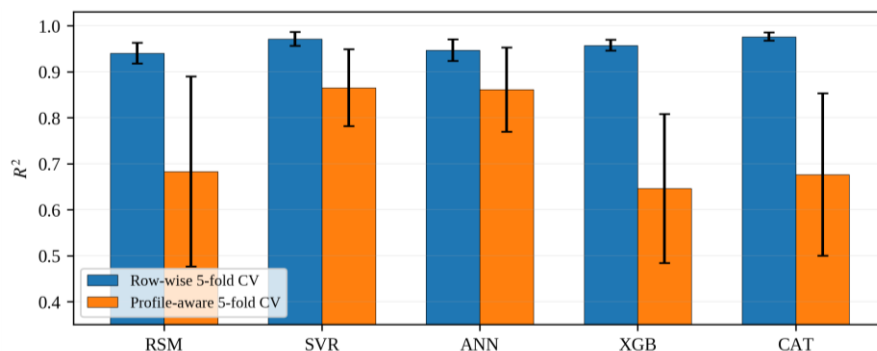


Figure 1. Row-wise and profile-aware five-fold validation estimates. Error bars represent the standard deviation among folds

The profile-aware robustness analysis changed the model ranking (Figure 1). CAT decreased from a row-wise CV mean R^2 of 0.976 to a grouped value of 0.676, and XGB to 0.646. SVR and ANN retained the strongest transfer, with grouped mean R^2 values of 0.865 and 0.861, respectively. This pattern is consistent with tree ensembles interpolating efficiently within represented kinetic series, while their apparent advantage diminishes when complete profiles are withheld. SVR was selected for screening because it achieved the highest grouped mean R^2 .

3.2. Profile-aware prediction and constrained screening

The pooled out-of-fold SVR predictions are presented in Figure 2a. Although larger than the row-wise errors, these values better represent prediction for an unseen profile. Dispersion was greatest at intermediate efficiency, where different one-factor kinetic blocks overlap; the grouped model should therefore be interpreted as an interpolation tool within the documented domain.

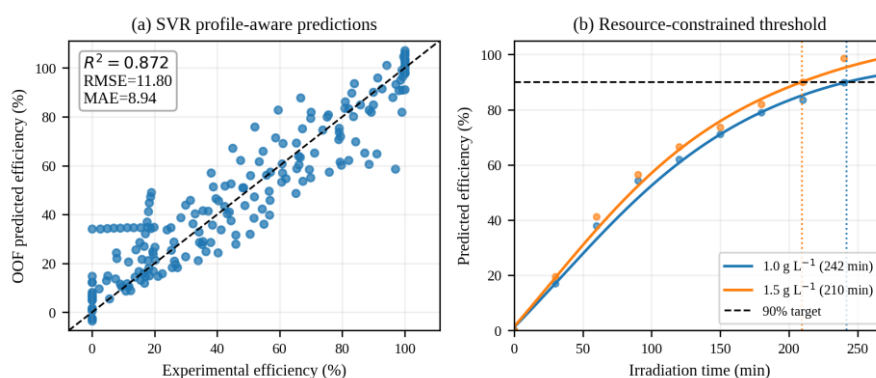


Figure 2. (a) Profile-aware SVR out-of-fold predictions; (b) discrete screening for the 90% efficiency threshold at two observed catalyst loadings (105 W, pH 5.45, HA = 0 mg L⁻¹, C₀ = 10 mg L⁻¹)

At 105 W, the 1.0 g L^{-1} condition reached the model-predicted 90% threshold at approximately 242 min, corresponding to a nominal irradiation-energy input of 423.5 Wh. At 1.5 g L^{-1} , the threshold occurred at 209.5 min and 366.6 Wh. Thus, the candidate reduced predicted irradiation time and nominal irradiation-energy input by 13.4%, while increasing catalyst loading by 50%. It represents a Pareto trade-off rather than an unconditional or experimentally confirmed optimum. The measured efficiency at 210 min and 1.5 g L^{-1} was 89.9%, supporting the local threshold estimate; nevertheless, independent experimental validation remains necessary.

Because the source experiments were primarily one-factor kinetic blocks, screening was limited to the observed factor arms. A designed experiment around $1.0\text{--}1.5 \text{ g L}^{-1}$ is required to validate interactions and refine the nominal irradiation-energy-catalyst-loading trade-off.

4. CONCLUSION

Profile-aware analysis of a public Pd–BiFeO₃ dataset showed that validation design materially changes estimated accuracy and model ranking. Row-wise splitting identified CAT as the leading model, whereas grouped separation favored SVR/ANN. The SVR surrogate identified a candidate condition that reduced predicted irradiation time and nominal irradiation energy input by 13.4%, at the cost of a 50% higher catalyst loading. Leakage-resistant validation, domain restrictions, and explicit resource trade-offs are therefore necessary for credible photocatalytic optimization.

ACKNOWLEDGEMENTS

The research presented in this paper was conducted with the financial support of the Ministry of Science, Technological Development, and Innovations of the Republic of Serbia, within the funding of the scientific research work at the University of Belgrade, Technical Faculty in Bor (contract number 451-03-34/2026-03/200131).

REFERENCES

- [1] V. Nedelkovski, M. Radovanović, S. Alagić, *Zaštita Materijala*, 65 (3) (2024) 524–533. doi:10.62638/ZasMat1007.
- [2] Z.H. Jaffari, A. Abbas, S.-M. Lam, S. Park, K. Chon, E.-S. Kim, K.H. Cho, J. Hazard. Mater., 442 (2023) 130031. doi:10.1016/j.jhazmat.2022.130031.
- [3] S. Das, S. Moon, R. Kaur, G. Sharma, P. Kumar, U. Lavrenčić Štanger, *Catal. Rev. Sci. Eng.*, 67 (3) (2025) 544–578. doi:10.1080/01614940.2024.2338131.
- [4] Z.H. Jaffari, Photocatalytic Performance Prediction, public data and code repository, 2022, github.com/ZeeshanHJ/Photocatalytic_Performance_Prediction (accessed 25 July 2026).
- [5] S. Kapoor, A. Narayanan, *Patterns*, 4 (9) (2023) 100804. doi:10.1016/j.patter.2023.100804.



The 57th International October Conference on Mining and Metallurgy

05 - 07 October 2026, Bor Lake, Serbia

<https://ioc.irmbor.co.rs>

- [6] C. Cortes, V. Vapnik, Mach. Learn., 20 (1995) 273–297. doi:10.1007/BF00994018.
- [7] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, Adv. Neural Inf. Process. Syst., 31 (2018) 6638–6648.
- [8] T. Chen, C. Guestrin, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
- [9] K. Miettinen, Nonlinear Multiobjective Optimization, Kluwer Academic Publishers, Boston, 1998. doi:10.1007/978-1-4615-5563-6.