Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
50(194):132-142, enero-marzo de 2026. doi: https://doi.org/10.18257/raccefyn.3245

Original article

Model selection for fungal laccase activity: shallow learners versus neural networks under nested cross-validation

Selección de un modelo para la actividad de la lacasa fúngica: modelos de aprendizaje superficial versus redes neuronales bajo validación cruzada anidada

Juan Carlos Montenegro[1], Mónica Miranda[2], Javier Torres[2], Rosa Elena Caballero[2,*]

[1] Universidad Tecnológica de Panamá, Ciudad de Panamá, Panamá
[2] Universidad Autónoma de Chiriquí, Provincia de Chiriquí, Panamá

Abstract

Due to their catalytic properties, fungal laccases are used in various technological fields, from bioremediation to biofuel production. Here, we revisited the model selection for extracellular laccase activity in *Trametes villosa* using a central composite design (CCD) with factors including temperature, initial pH, and inoculum volume. To address concerns about overfitting and evaluation leakage in small-N RSM studies, we adopted a strict nested cross-validation (CV) protocol (outer 5-fold for generalization and inner 3-fold for tuning) and compared three shallow learners: quadratic ridge regression, RBF-kernel support-vector regression (SVR), and random forests, against two compact multilayer perceptrons (MLPs). To preserve physical plausibility, we modeled the response as $\log(y + \varepsilon)$ with $\varepsilon = 0.1$, and reported RMSE and R^2 from the outer 5-fold CV on the back-transformed scale. Across outer folds, SVR(RBF) achieved the lowest RMSE with the smallest fold-to-fold variance; the random forest were competitive; MLPs were more variable, and the ridge underfitted. SHAP analyses for the best model highlighted temperature as the dominant driver, with pH and inoculum showing secondary, non-monotonic contributions. Our results indicate that shallow nonlinear methods generalize best on small RSM datasets and should be preferred for early-stage process optimization.

Keywords: Central composite design; Fermentation; Machine learning; White-rot fungi.

Resumen

Debido a sus propiedades catalíticas, las lacasas fúngicas tienen aplicación en diversos campos tecnológicos, desde la biorremediación hasta la producción de biocombustibles. Aquí reexaminamos la selección de modelos para la actividad de la lacasa extracelular en *Trametes villosa* mediante un diseño compuesto central (CCD), considerando factores como la temperatura, el pH inicial y el volumen de inóculo. Para abordar las preocupaciones sobre el sobreajuste y la "fuga" en la evaluación en estudios RSM de tamaño muestral pequeño (N pequeño), adoptamos un protocolo estricto de validación cruzada (CV) anidada (externa de 5 pliegues para la generalización e interna de 3 pliegues para el ajuste de hiperparámetros) y comparamos tres modelos de aprendizaje superficial: regresión de cresta cuadrática, regresión de vectores de soporte (SVR) de núcleo RBF y bosques aleatorios con dos perceptrones multicapa compactos (MLP). Para preservar la plausibilidad física, modelamos la respuesta como $\log(y + \varepsilon)$ con $\varepsilon = 0{,}1$ y reportamos el RMSE y el R^2 del CV externo de 5 pliegues en la escala retrotransformada. En los pliegues externos, la SVR (RBF) obtuvo el menor RMSE y la menor varianza entre pliegues; el bosque aleatorio fue competitivo; los MLP mostraron mayor variabilidad y la cresta se ajustó insuficientemente. Los análisis SHAP para el mejor modelo destacaron la temperatura como el principal factor determinante, y el pH y el inóculo como contribuciones secundarias, no monótonas. Los resultados indican que los métodos no lineales superficiales se generalizan mejor en conjuntos de datos RSM pequeños y deberían preferirse para la optimización de procesos en etapas tempranas.

Palabras clave: Aprendizaje automático; Diseño central compuesto; Fermentación; Hongos de podredumbre blanca.

Citation: Montenegro JC, *et al.* Model selection for fungal laccase activity: shallow learners versus neural networks under nested cross-validation. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales. 50(194):132-142, enero-marzo de 2026. doi: https://doi.org/10.18257/raccefyn.3245

Editor: María Mercedes Zambrano

***Corresponding autor:**
Rosa Elena Caballero;
rosa.caballero@unachi.ac.pa

Received: July 13, 2025
Accepted: November 19, 2025
Published on line: February 6, 2026

Introduction

Fungal laccases are multicopper blue oxidases that reduce molecular oxygen to water by means of a one-electron oxidation (**Wadhwa *et al.***, 2023). In nature, these enzymes participate in fungal defense mechanisms; they also play a role in the degradation of wood structural components, litter decomposition, sporulation, and pigmentation (**Janusz *et al.*,** 2020). Due to these capabilities and reaction mechanisms, laccases are recognized as green catalysts (**Agrawal *et al.***, 2018; **Mayolo-Deloisa *et al.***, 2020) with high potential in circular approaches. For instance, laccases can degrade lignocellulosic biomass from agro-industrial activities, rendering biomaterials that can be used as soil amenders, animal food supplements, or biosorbents (**Loi *et al.***, 2021; **Mondal *et al.***, 2023). They have also been used to treat effluents from the paper, textile, pharmaceutical, and alcohol industries (**Paraschiv *et al.***, 2022).

The first step in laccase technology development is enzyme production. The adjustment of culture conditions, for example, the type and amount of carbon and nitrogen sources, the presence of cofactors and inducers, the initial pH, and the temperature, play a significant role in enzyme production (**Dhakar & Pandey**, 2013; **Karp *et al.***, 2015). This information is key to scaling-up and application at an industrial scale.

Response surface methodologies (RSM) have been widely used to study the influence of fermentation conditions on one or more outputs (**Aragao *et al.***, 2020; **Senthivelan *et al.***, 2019). Artificial intelligence-based machine learning techniques (AI-ML) have also been introduced to help in data modeling, prediction, and classification (**Cheng *et al.***, 2022; **Chentamara *et al.***, 2022; **da Silva Pereira *et al.***, 2021; **de Meneses *et al.***, 2023; **Wainaina & Taherzadeh**, 2023).

Several authors have compared statistical techniques and AI-ML approaches. For instance, **Dourado Fernandes *et al.*** (2020) applied an artificial neural network coupled with a genetic algorithm (ANN-GA) and RSM to predict Reactive Black 5 decolorization by crude enzymes from *Pleurotus sajor-caju*. Similarly, **Vilar *et al.*** (2021) optimized enzyme production from *Pleurotus sajor-caju* using RSM and ANN-GA coupled models.

The catalytic efficiency and optimal conditions for enzyme catalysis have also been explored through computational methods. **Cao *et al.*** (2024) constructed a machine learning model with phosphatase (EC 3.1.3.X) using amino acid frequency and protein molecular weight information as features. They applied the K-nearest neighbors regression algorithm to predict optimal temperature. **Alazmi** (2024) evaluated catalytic efficiency prediction with conventional neural networks and XGBoost, while optimum pH was addressed by **Gado *et al.*** (2025) and **Zhang *et al.*** (2025).

In a previous experiment, we used a central composite design (CCD) to model and optimize fungal laccase activity in a native strain of the white-rot fungus *T. villosa*, focusing on temperature, pH, and inoculum volume as inputs. Despite trying standard response transformations, the statistical fit was limited. Here, we revisited the same dataset to perform principled model selection under a nested cross-validation protocol that avoids data leakage in small designs. We compared shallow learners (quadratic ridge regression, support-vector regression with an RBF kernel, and random forests) with compact neural networks, and used SHAP to interpret the best model predictions in terms of process variables.

Materials and methods

Organism and fermentation conditions

Trametes villosa (strain LG72) was provided by the Universidad Autónoma de Chiriquí (UCH) Herbarium. It was kept by monthly transfers in Petri dishes with an agar basal medium (**Gutiérrez *et al.***, 2015; **Moore *et al.***, 2020) at 26-28°C, alternating 12-hour periods of light and darkness. The basal medium composition in g/L was as follows: KCl, $MgSO_4 \cdot 7H_2O$ (0.5 of each); $FeSO_4 \cdot 7H_2O$, $ZnSO_4$, $CuSO_4 \cdot 5H_2O$ (0.01 of each), KH_2PO_4 1.0, thiamine 0.001, asparagine 2.0, glucose 10.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
50(194):132-142, enero-marzo de 2026. doi: https://doi.org/10.18257/raccefyn.3245

Using 250 mL cell culture flasks, we adjusted 50 mL of the basal fermentation medium to different pH values and temperatures, and we inoculated them with different volumes of a hyphal suspension according to the CCD matrix (**Table 1**). This suspension was prepared as reported in **Caballero *et al.*** (2024). We reached a 120-rpm agitation speed in a shaker incubator (Shel Lab SI4, Sheldon Manufacturing Inc., USA). After 10 days of fermentation, we vacuum filtered the contents of each flask and kept the filtrate at 7 °C before enzymatic analysis.

Enzymatic analysis

To determine laccase activity (U/L), the filtrates were stabilized at room temperature and mixed with 50 μL of siringaldazine 0.22 mM (Sigma Aldrich, USA), at 26°C and 1 mL of 100 mM phosphate buffer pH 6.5. The change in absorbance at 530 nm (ε = 65,000 $M^{-1}cm^{-1}$) in two minutes was followed with a Thermo Scientific BiomateTM 6 spectrophotometer. One unit of laccase activity was defined as the amount of enzyme needed to oxidize 1 µmol of syringaldazine per mL per min (**Junior *et al.***, 2020; **Palmieri *et al.***, 1998). Laccase activity (**Table 1**) was reported as the average value from three replicates.

Dataset

The dataset obtained from the CCD matrix consisted of 48 randomized experimental runs with six replicates at central points and three at axial points. Three factors were evaluated at three levels (low, medium, and high) in each of them: pH (5, 6, 7), temperature (28, 32, 36°C), and inoculum volume (10, 20, 30 mL). The enzymatic activity expressed in units per liter (U/L) was the output (**Table 1**)**.** The data were imported from a CSV file for analysis.

Model development

Response transformation and positivity constraint. Given that enzyme activity cannot be negative and often varies widely, we first transformed the response so the model would avoid impossible negative predictions and treat low and high values more evenly.

To preserve physical plausibility (non-negative activity) and stabilize variance, the response was modeled on a log-shifted scale: $y' = log(y + \varepsilon)$ with $\varepsilon = 0.1$. This choice guaranteed that back-transformed predictions remained non-negative and mitigated the undue influence of high-activity outliers. Unless otherwise stated, all performance metrics were reported on the original (back-transformed) scale, i.e., $\hat{y} = \max\{\exp(\hat{y}') - \varepsilon, 0\}$.

Candidate models and hyperparameters. We compared three shallow learners and two compact neural networks within reproducible scikit-learn pipelines:

- Ridge (poly2): PolynomialFeatures (degree=2, no bias) → StandardScaler → Ridge
  Hyperparameters: $\alpha \in \{0.1, 1.0, 10.0\}$
- SVR (RBF): StandardScaler → Support Vector Regression (RBF kernel).
  Hyperparameters: $C \in \{1, 10\}, \gamma \in \{\text{“scale”}, 0.1\}, \epsilon = 0.1$
- Random forest: Passthrough preprocessing → RandomForestRegressor (random_state=0)
  Hyperparameters: $n_{estimators} \in \{300, 600\}$, max _depth $\in$ None, 6,12}, min _samples_leaf $\in \{1,2,3\}$
- MLP small (16, 8) and MLP medium (32, 16): StandardScaler → MLPRegressor (ReLU, early_stopping=True, learning_rate_init=10^{-3}, max_iter=3000,random_state=0)
  Hyperparameters: $\alpha \in \{10^{-4}, 10^{-3}\}$

All feature engineering (polynomial expansion and scaling) was encapsulated inside each pipeline to avoid data leakage during cross-validation.

Nested cross-validation protocol. Generalization was assessed with nested cross-validation. The outer loop used KFold (5 splits, shuffle=True, random_state=42) to produce unbiased test folds. Within each outer training set, the inner loop used KFold (3 splits, shuffle=True, random_state=123) and GridSearchCV (scoring = negative MSE on

Table 1. Central composite design (CCD) dataset used for model development (n = 48). Response: extracellular laccase activity (U/L). Factors and units: Factor 1 = Temperature (°C), Factor 2 = Initial pH, Factor 3 = Inoculum volume (mL)

Run	Temperature (°C)	pH	Inoculum volume (mL)	Laccase activity (U/L)
1	32	6	20	19.24
2	32	6	10	10.25
3	32	6	30	1.29
4	32	5	20	11.35
5	32	6	20	17.29
6	36	5	10	11.07
7	32	6	30	1.19
8	36	7	30	0.25
9	28	7	30	15.07
10	36	7	10	0.03
11	32	6	20	16.18
12	28	5	10	0.13
13	28	7	10	2.72
14	32	6	20	19.14
15	28	5	30	0.28
16	28	6	20	3.72
17	28	6	20	3.63
18	32	6	10	9.94
19	36	5	10	15.74
20	36	6	20	0.16
21	28	5	30	0.79
22	36	7	30	0.22
23	36	7	10	0.03
24	32	6	20	16.04
25	36	5	30	7.66
26	36	6	20	0.31
27	32	7	20	1.19
28	32	7	20	3.41
29	36	5	30	6.69
30	32	5	20	12.55
31	28	7	10	5.36
32	28	6	20	4.04
33	28	7	10	2.81
34	36	7	10	0.06
35	36	6	20	0.19
36	32	7	20	2.4
37	32	5	20	14.7
38	32	6	10	11.35
39	36	7	30	0.25
40	36	5	10	13.22
41	28	5	30	0.6
42	28	7	30	13.18
43	28	5	10	0.06
44	32	6	20	17.76
45	32	6	30	0.94
46	36	5	30	4.37
47	28	5	10	0.1
48	28	7	30	14.95

Table 1 shows 48 randomized experiments with six replicates at central points.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
50(194):132-142, enero-marzo de 2026. doi: https://doi.org/10.18257/raccefyn.3245

the transformed scale) to select hyperparameters. The selected model was refitted on the entire outer-train partition and evaluated on the held-out outer-test partition. Outer-fold predictions were back-transformed to the original scale before computing metrics.

Model interpretability (pre-specified). After model selection, the best model was refitted on the full dataset and interpreted with SHAP (**Lundberg & Lee**, 2017) on the back-transformed scale. For tree models, we used TreeExplainer, and for non-tree models (SVR, Ridge, MLP), we used KernelExplainer on the full pipeline with a small background sample ($n \leq 30$) and fixed seeds. Outputs include a beeswarm plot and a mean-|SHAP| importance plot; numerical results appear in the next section.

Implementation details and outputs. Analyses were performed in *Python* using scikit-learn (**Pedregosa *et al.***, 2011; **Van Rossum & Drake**, 2009). The driver script (compare_shallow_vs_nn_rf_shap.py) (i) loaded and cleaned the CSV (decimal, non-numeric stripping), (ii) inferred the response column from common keywords (e.g., "activity", "U/L") and selected numeric features, (iii) ran nested CV for all models, and (iv) exported:

- metrics.csv: outer-CV summary per model RMSE_mean, RMSE_SD, $Cl_{95}, R^2 - mean, R^2 - SD, Cl_{95}$
- predictions.csv: row-level outer-CV predictions for all models,
- rmse_bar.png: RMSE (mean ± SD) bar chart across models,
- parity_best.png: observed vs. predicted plot for the globally best model,
- shap_summary.png, shap_bar.png: SHAP beeswarm and mean |SHAP| plots for the best refit model.

Evaluation metrics. All performance metrics were computed on the original scale after back-transforming predictions from the modeling scale $y' = log(y + \varepsilon)$ with $\varepsilon = 0.1$; specifically, $\hat{y} = max\{exp(\hat{y}') - \varepsilon, 0\}$ (U/L). The primary metric was the root-mean-squared error (RMSE),

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

and the secondary metric was the coefficient of determination on the original scale,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \hat{y})^2} \qquad (2).$$

Under nested cross-validation (outer 5-fold, inner 3-fold), we reported for each model the outer-fold mean and RMSE and R^2 standard deviation (SD), together with bootstrap 95% confidence intervals on fold means (1,000 bootstrap resamples). Model selection favored the lowest outer-CV RMSE; ties were broken by lower variance across folds and a higher R^2. Parity plots using concatenated outer-fold predictions were used as a visual calibration check.

Results

Overall model comparison (nested CV)

Table 2 summarizes performance across models. SVR(RBF) achieved the lowest RMSE and smallest fold-to-fold variance (RMSE = 1.321 ± 0.231 U/L; R^2 = 0.955 ± 0.016). Random forest was competitive (RMSE = 1.908 ± 0.371 U/L; R^2 = 0.907 ± 0.038). Both MLPs (small/medium) exhibited higher variability under small-N, and ridge (poly2) underfitted (RMSE =6.366±1.294 U/L; negative R^2). **Figure 1** shows the RMSE mean ± SD per model family.

Generalization quality (parity)

The parity plot for the best model (SVR(RBF)) showed close alignment to the identity line across the activity range, with modest dispersion at high activities (**Figure 2**), indicating good calibration on the original scale.

Table 2. Outer 5-fold nested-CV performance on the back-transformed scale after log(y+ε) with ε = 0.1. Values are means across outer folds with SD and bootstrap 95% CIs.

Model	Category	rmse_mean	rmse_sd	rmse_ci95_lo	rmse_ci95_hi	r^2_mean	r^2_sd	r^2_ci95_lo	r^2_ci95_hi
MLP medium (32,16) [nn]	nn	2.618	1.712	1.492	4.34	0.7688	0.2906	0.4721	0.9454
MLP small (16,8) [nn]	nn	5.41	3.445	2.405	8.334	-0.0133	0.8407	-0.7488	0.7222
SVR(RBF) [shallow]	**shallow**	**1.321**	**0.2312**	**1.098**	**1.509**	**0.9551**	**0.01632**	**0.9416**	**0.9701**
Random Forest [shallow]	shallow	1.908	0.3713	1.617	2.257	0.9073	0.03817	0.8694	0.9346
Ridge(poly2) [shallow]	shallow	6.366	1.294	5.066	7.208	-0.02666	0.3395	-0.2634	0.2971

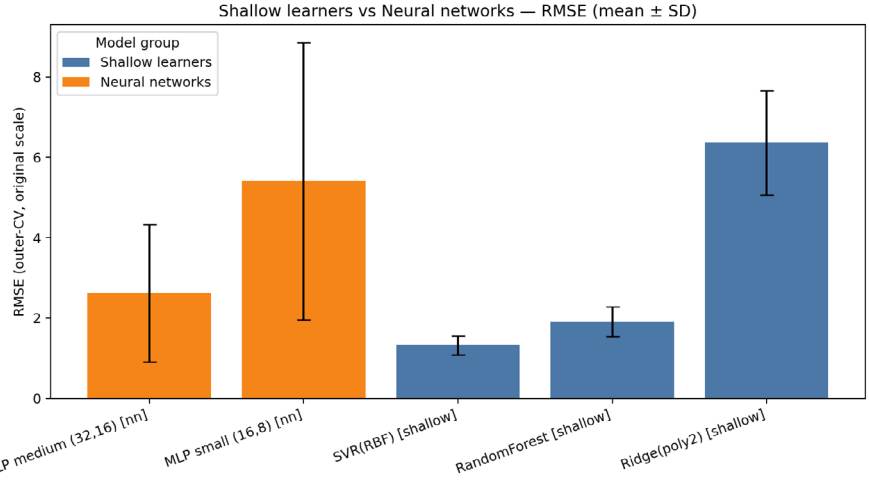

Figure 1. Outer 5-fold nested-CV performance (RMSE, mean ± SD) on the original scale after back-transforming predictions from log (y+ε) with ε = 0.1. Error bars denote across-fold SD; inner 3-fold CV was used for hyperparameter tuning. Models shown (left→right): MLP medium (32, 16), MLP small (16, 8), SVR (RBF), random forest, and ridge (poly2). Lower is better. Colors distinguish shallow learners (blue) and neural networks (orange).

Robustness and sensitivity checks

The model ranking did not change when varying the positivity shift within ε ∈ [0.05, 0.2] and modestly widening the SVR *C* and RF max_depth grids, which supported that the observed advantage of shallow nonlinear methods was not driven by a particular configuration.

Error analysis

Outer-CV residuals were approximately homoscedastic over most of the range, with slightly larger absolute deviations at higher activities. No single factor level dominated the largest residuals, consistent with interaction effects captured by SVR (RF).

Model interpretability (SHAP)

Temperature was the dominant driver of extracellular laccase activity, with the initial pH and inoculum volume contributing secondarily and non-monotonically. **Figure 3a** (beeswarm) and **b** (mean-|SHAP|) summarize these effects, indicating temperature bands with the strongest marginal gains and narrower pH windows.

Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales.
50(194):132-142, enero-marzo de 2026. doi: https://doi.org/10.18257/raccefyn.3245

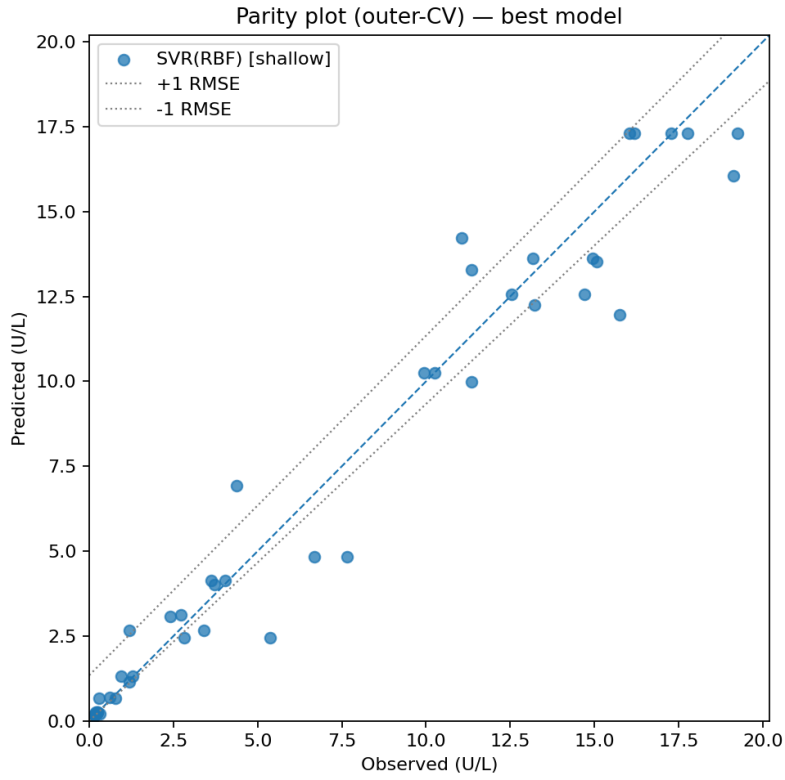

Figure 2. Parity plot (observed vs. predicted) on outer test folds for the best-performing model SVR (RBF). Predictions are back-transformed from log (y+ε) with ε = 0.1; units: U/L. The dashed line denotes perfect agreement; points aggregate all outer folds. Close clustering around the identity line indicates good calibration, with a modest spread at higher activities. Dashed line: identity; dotted bands: ± 1 RMSE

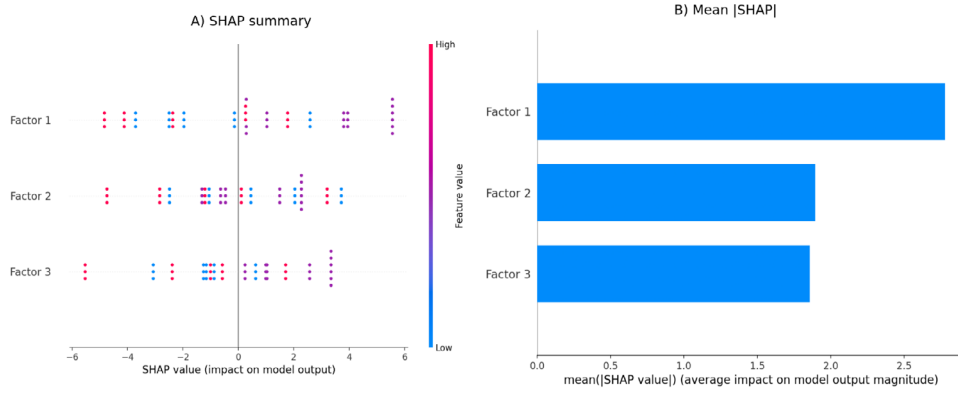

Figure 3. a) SHAP beeswarm for the best refit model SVR (RBF) explaining back-transformed predictions. Each point is a sample; horizontal position shows the feature's marginal impact on the prediction (SHAP value), and color encodes the original feature value (low→high). Features are ordered by overall importance; Factor 1 (temperature) dominates, while Factor 2 (initial pH) and Factor 3 (inoculum volume) exhibit secondary, non-monotonic effects. **b)** Feature importance by mean |SHAP| for the best refit model SVR (RBF) explaining back-transformed predictions. Bars report mean absolute SHAP values (U/L); right-hand labels indicate each feature's share of total importance. Factor 1 (temperature) dominates, followed by Factor 2 (initial pH) and Factor 3 (inoculum volume).

Process-level interpretation. SHAP patterns suggested temperature bands where marginal gains were the strongest, beyond which returns diminished; pH showed windowed effects consistent with enzyme ionization states; inoculum effects were present but less pronounced and context dependent. Practically, this argues for tight temperature control as the primary lever, complemented by pH windowing and moderate inoculum adjustments.

Discussion

Practical implications for early-stage process optimization

Along a strict nested cross-validation protocol, shallow nonlinear models provided the most reliable accuracy–stability trade-off for this small, CCD-style RSM dataset. Using pipelines that confined preprocessing within each fold, and reporting performance strictly from the outer 5-fold CV, SVR(RBF) consistently achieved the lowest RMSE with the smallest fold-to-fold variance, while random forests were competitive, multilayer perceptrons (MLPs) exhibited higher variance across folds, and quadratic ridge underfitted the signal. All performance metrics (RMSE and R^2) were computed from outer-fold predictions on the back-transformed scale after modeling $\log(y + \varepsilon)$, which ensured that results reflected the expected generalization rather than an in-sample fit, and remained in physically meaningful units.

Methodologically, the combination of nested CV and fold-internal preprocessing mitigated data leakage, an often-overlooked source of optimism in small-N studies. Reporting outer-fold means and dispersion, together with visualization via parity plots (with ±1 RMSE bands), calibrated expectations for typical error, and highlighted heterogeneity across folds. This design choice prioritized estimation accuracy and external validity while keeping the multiple-comparison burden of tuning under control.

Model interpretability further supported these conclusions. SHAP analyses of the best model identified temperature as the dominant driver of extracellular laccase activity, with initial pH and inoculum volume contributing in secondary, non-monotonic ways. These patterns are consistent with plausible bioprocess behavior (curvature and interactions without excessive parameterization), reinforcing the face validity of the selected shallow models under data scarcity.

Practical implications for early-stage process optimization are direct. In small factorial or CCD experiments where sample size and factor levels are limited, shallow nonlinear learners (e.g., SVR with RBF kernels and random forests) should be the first-line baselines and screening tools, as they capture key interactions and curvature with stable generalization under nested CV, allowing practitioners to prioritize factor regions and hypotheses before committing resources to larger neural architectures or global surrogates that typically require more data to control variance.

Limitations and future work

Limitations include the inherently small-N setting and deliberately modest hyperparameter grids to reduce multiplicity within nested CV. Future work should (i) investigate shape-constrained or monotonic models to encode biochemical priors, (ii) adopt Bayesian optimization or sequential experimental design to target informative regions of the factor space, and (iii) perform external validation across independent batches and strains to assess portability. Extensions to uncertainty-aware modeling (e.g., quantile regression or conformal prediction) and careful treatment of measurement noise and potential heteroscedasticity would further strengthen prospective decision-making in process development.

Conclusions

Using a CCD dataset for *Trametes villosa* laccase activity and a strict nested CV protocol, shallow nonlinear learners delivered the best balance of bias–variance and robustness at small N. SVR (RBF) provided the strongest generalization with low variance, random forest was a reliable runner-up, compact MLPs were less stable, and ridge (poly2) underfitted the

observed nonlinearities. SHAP analyses identified temperature as the dominant driver, with initial pH and inoculum volume exerting secondary, non-monotonic effects. For early-stage process optimization, we therefore recommend starting with SVR (RBF) baselines under nested CV, tightening temperature control and exploring pH windows, and considering larger neural models only when data volume and experimental coverage grow.

Authors' contribution

JCM: data curation, analysis, and interpretation, authored the article; MM: laboratory work, data collection, helped with the original draft; **JT:** data analysis, helped with the original draft; **REC:** project idea and work team supervision, reviewed the original draft and the last version of the manuscript.

Conflicts of interest

The authors have no competing interests to declare.

Acknowledgments

The authors thank Dr. Tina Hoffman, curator at the Universidad Autónoma de Chiriquí (UCH) Herbarium, for providing the *Trametes villosa* strain, and Mrs. Dayra Icaza, technician at the Microbiology Laboratory, Universidad Autónoma de Chiriquí, for her support with laboratory equipment. We also thank Professor Víctor Jiménez for his help in data analysis.

Data and code availability

All analysis code and plotting scripts used in this study are publicly available at GitHub (https://github.com/juanky2797/fungal-laccase-activity-model). The repository contains the driver script compare_shallow_vs_nn_rf_shap.py, which (i) loads the CSV dataset, (ii) performs nested cross-validation (outer=5, inner=3) for the five model families, (iii) reports metrics on the back-transformed scale after log(y+ε) with ε=0.1, and (iv) generates all figures and artifacts used in the manuscript. Running the script produces the following outputs alongside the input CSV:

- metrics.csv (model-level outer-CV summary),
- predictions.csv (row-level outer-CV predictions),
- rmse_bar.png (**Figure 1**), parity_best.png **(Figure 2**),
- shap_summary.png and shap_importance.png (**Figure 3**).

References

Agrawal, K., Chaturvedi, V., Verma, P. (2018). Fungal laccase discovered but yet undiscovered. *Bioresources & Bioprocessing*, *5*, 4. https://doi.org/10.1186/s40643-018-0190-z

Alazmi, M. (2024). Enzyme catalytic efficiency prediction: employing convolutional neural networks and XGBoost. *Frontiers in Artificial Intelligence*, *7*, 1446063. https://doi.org/10.3389/frai.2024.1446063

Aragao, M.S., Menezes, D.B., Ramos, L.C., Oliveira, H.S., Bharagava, R.N., Ferreira, L.F.R., Teixeira, J.A., Ruzene, D., Silva, D.P. (2020). Mycoremediation of vinasse by surface response methodology and preliminary studies in air-lift bioreactors. *Chemosphere*, *244*(2), 125432. https://doi.org/10.1016/j.chemosphere.2019.125432

Caballero, R. E., Jiménez, V., Miranda, M., Rovira, D., Branda, G., de Pérez, J. R. C. (2024). An Optimal Culture Medium for Laccase Production and Sugar Cane Vinasse Biotreatment with *Trametes villosa* Using Plackett-Burman and Central Composite Designs. *Applied Environmental Research*, *46*(1), 001. https://doi.org/10.35762/AER.2024001

Cao, Y., Qiu, B., Ning ,X., Fan, L., Qin, Y., Yu, D., Yang, C., Ma, H., Liao, X., You, C. (2024). Enhancing Machine-Learning Prediction of Enzyme Catalytic Temperature Optima Through Amino Acid Conservation Analysis. *International Journal of Molecular Science*, *25*(11), 6252. https://doi.org/10.3390/ijms25116252

Cheng, Y, Bi, X., Xu, Y., Liu, Y., Li, J., Du, G., Lv, X., Liu, L. (2022). Artificial intelligence technologies in bioprocess: Opportunities and challenges. *Bioresource Technology, 369*, 128451. https://doi.org/10.1016/j.biortech.2022.128451

Chenthamara, D., Sivaramakrishnan, M., Ramakrishnan, S., Esakkimuthu, S., Ko, R., Subrmaniam, S. (2022). Improved laccase production from *Pleurotus floridanus* using deoiled microalgal biomass: statistical and hybrid swarm-based neural networks modeling approach. *3 Biotech, 12* (12), 346. https://doi.org/10.1007/s13205-022-03404-y

da Silva Pereira, A., Pinheiro, Á.D.T., Rocha, M.V.P., Gonçalves, L.R.B., Cartaxo, S.J.M. (2021). Hybrid neural network modeling and particle swarm optimization for improved ethanol production from cashew apple juice. *Bioprocess Biosystems Engineering*, *44*(2), 329-342. https://doi.org/10.1007/s00449-020-02445-y

de Menezes, L.H.S., Pimentel, A.B., Oliveira, P.C., Tavares, I.M., Ruiz, H., Irfan, M., Bilal, M., das Chagas, T.P., da Silva, R.G.P., Salay, L.C., de Oliveira, J.R., Franco, M. (2023). The Application of Chemometric Methods in the Production of Enzymes Through Solid State Fermentation Uses the Artificial Neural Network—a Review. *Bioenergy Research*, *16*, 279–288. https://doi.org/10.1007/s12155-022-10462-w

Dhakar, K. & Pandey, A. (2013). Laccase Production from a Temperature and pH Tolerant Fungal Strain of *Trametes hirsuta* (MTCC 11397). *Enzyme Research*, *2013,* 869062. https://doi.org/10.1155/2013/869062

Dourado Fernandes, C., Nascimento, V., Menezes, D., Vilar, D., Torres, N., Leite, M., Veja-Baudrit, J., Bilal, M., Iqbal, H., Bharagava, R., Egues, S., Ferreira, L.F. (2020). Fungal biosynthesis of lignin-modifying enzymes from pulp wash and *Luffa cylindrica* for azo dye RB5 biodecolorization using modeling by response surface methodology and artificial neural network. *Journal of Hazardous Materials, 399,* 123094. https://doi.org/10.1016/j.jhazmat.2020.123094

Gado, J.E., Knotts, M., Shaw, A.Y., Marks, D., Gauthier, N.P., Sander, C., Beckham, G.T. (2025). Machine learning prediction of enzyme optimum pH. *Nature Machine Intelligence, 7,* 716-729. https://doi.org/10.1038/s42256-025-01026-6

Gutiérrez-Soto, G., Medina-González, G., Treviño-Ramírez, J., Hernández-Luna C. (2015). Native macro fungi that produce lignin-modifying enzymes, cellulases and xylanases with potential biotechnological applications. *BioResources, 10*(4), 6676-6689. https://doi.org/10.15376/biores.10.4.6676-6689

Janusz, G., Pawlik, A., Świderska, U., Polak, J., Justyna, S., Jarosz-Wilkolazka, A., Paszczynski, A. (2020). Laccase Properties, Physiological Functions, and Evolution. *International Journal of Molecular Sciences, 21,* 966. https://doi.org/10.3390/ijms21030966

Junior, J.A., Vieira, Y.A., Cruz, I.A., Vilar, D., Aguiar, M.M., Torres, N.H., Bharagava, R.N., Lima, A.S., de Souza, R.L., Ferreira, L.F.R. 2020. Sequential degradation of raw vinasse by a laccase enzyme producing fungus *Pleurotus sajor-caju* and its ATPS purification. *Biotechnology Reports*, *25*, e00411.

Karp, S., Faraco, V., Amore, A., Letti, L., Thomaz-Soccol, V., Soccol, C. (2015). Statistical Optimization of Laccase Production and Delignification of Sugarcane Bagasse by *Pleurotus ostreatus* in Solid-State Fermentation. *BioMed Research International, 2015*, 1-8. https://doi.org/10.1155/2015/181204

Loi, M., Glazunova, O., Fedorova, T., Logrieco, A.F., Mulè, G. (2021). Fungal Laccases: The Forefront of Enzymes for Sustainability. *Journal of Fungi, 7,* 1048. https://doi.org/10.3390/jof7121048

Lundberg, S. M., Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In U. von Luxburg. (Ed.), *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 4768-4777).Curran Associates Inc.

Mayolo-Deloisa K, González-González M., Rito-Palomares M. (2020). Laccases in Food Industry: Bioprocessing, Potential Industrial and Biotechnological Applications. *Frontiers in Bioengineering and Biotechnology, 8*, 222. https://doi.org/10.3389/fbioe.2020.00222

Mondal, P., Galodha, A., Verma, V., Singh, V., Show, P. Awasthi, M., Lall, B., Anees, S., Pollmann, K., Jain, R. (2023). Review on machine learning-based bioprocess optimization, monitoring, and control systems. *Bioresource Technology, 370*, 128523. https://doi.org/10.1016/j.biortech.2022.128523

Moore, D., Robson, G., Trinci, A. (2020). *21st Century guidebook to fungi*. Cambridge University Press. https://doi.org/10.1017/9781108776387

Palmieri, G., Giardina, P., Bianco, C., Scaloni, A., Capasso, A., Sannia, G. (1998). A Novel white laccase from *Pleurotus ostreatus*. *The Journal of Biological Chemistry, 272*(50), 31301-31307. https://dx.doi.org/10.1074/jbc.272.50.31301

Paraschiv, G., Ferdes, M., Ionescu, M., Moiceanu, G., Zabava, B.S., Dinca, M.N. (2022). Laccases—Versatile Enzymes Used to Reduce Environmental Pollution. *Energies, 15*, 1835. https://doi.org/10.3390/en15051835

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.

Senthivelan, T., Kanagaraj, J., Panda, R.C., Narayani, T. (2019). Screening, and production of a potential extracellular fungal laccase from *Penicillium chrysogenum*: Media optimization by response surface methodology (RSM) and central composite rotatable design (CCRD). *Biotechnology Reports*, *23*, 30-44. https://doi.org/10.1016/j.btre.2019.e00344

Van Rossum, G. & Drake, F. L. (2009). Python 3 Reference Manual. Python Software Foundation. CreateSpace.

Vilar, D., Dourado Fernandes, C. Nascimento, V., Torres, N., Leite, M., Bharagava, R., Bilal, M., Salazar, G., Eguiluz, K., Ferreira, L.F. (2021). Hyper-production optimization of fungal oxidative green enzymes using citrus low-cost byproduct. *Journal of Environmental Chemical Engineering*, 9, 105013. https://doi.org/10.1016/j.jece.2020.105013

Wadhwa, H., Singh, R., Chopra, C. (2023). Occurrence and Applications of Fungal Laccases: A Comprehensive Biotechnological Review. *Biological Forum – An International Journal, 15*(4), 463-469(2023).

Wainaina, S. & Taherzadeh, M. (2023). Automation and artificial intelligence in filamentous fungi-based bioprocesses: A review. *Bioresource Technology, 369*, 128421. https://doi.org/10.1016/j.biortech.2022.128421

Zhang, L., Luo, K., Zhou, Z., Yu, Y., Jiang, F., Banghao, W., Li, M., Hong, L. (2025). A Deep Retrieval-Enhanced Meta-Learning Framework for Enzyme Optimum pH Prediction. *Journal of Chemical Information and Modeling*, *65*(7), 3761-3770. https://doi.org/10.1021/acs.jcim.4c02291