processes

MDPI

Article

Replicate-Aware Model Selection and Surrogate-Based Optimization for Predicting and Maximizing Phenolic Removal from Sugarcane Vinasse

Nicomedes Ariel Jaramillo [1], María Córdoba [2], Mónica Miranda [3], Juan Carlos Montenegro [4] and Rosa Elena Caballero [3,*]

[1] Programa de Maestría en Investigación, Vicerrectoría de Investigación y Posgrado, Universidad Autónoma de Chiriquí, David 0427, Panama; nicomedes.jaramillo@unachi.ac.pa
[2] Escuela de Química, Facultad de Ciencias Naturales y Exactas, Universidad Autónoma de Chiriquí, David 0427, Panama; maria.cordoba2@unachi.ac.pa
[3] Centro Especializado en Investigaciones de Química Inorgánica y Afines, Universidad Autónoma de Chiriquí, David 0427, Panama; monica.miranda1@unachi.ac.pa
[4] Facultad de Ingeniería de Sistemas Computacionales, Universidad Tecnológica de Panamá, Panama City P.O. Box 0819-07289, Panama; juan.montenegro1@utp.ac.pa
* Correspondence: rosa.caballero@unachi.ac.pa

Abstract

Machine learning is increasingly used to predict pollutant removal in adsorption studies, but models are often fitted to small, replicated experimental datasets and evaluated with a single random train/test split. We examined how the validation design affects model selection and surrogate-based optimization for removing phenolic compounds (quantified as tannins) from sugarcane vinasse using citric-acid-modified sugarcane bagasse, based on 315 measurements (105 conditions in triplicate) spanning temperature, contact time, and vinasse dilution. Eight regressors, from linear and quadratic models to kernel methods, tree ensembles, Gaussian processes, and neural networks, were tuned and evaluated by repeated nested cross-validation that kept all replicates of a condition together, and their errors were compared with the replicate (pure-error) standard deviation of 3.56% removal. Under this replicate-aware validation, all non-linear models performed similarly (RMSE 3.89–4.10%, R^2 0.83–0.85), whereas the linear model was clearly worse (RMSE 5.06%). A naive random split increased R^2 by only about 0.01 (mean ΔR^2 0.005–0.013), because, with replicate scatter of this size, memorizing replicates gives no advantage over a smooth model. Holding out entire dilution levels, however, separated the models sharply: the quadratic model remained the most accurate (RMSE 4.58%), whereas random forest and gradient boosting failed ($R^2 \leq 0.13$). Permutation and SHAP analyses identified dilution as the dominant variable. Used as the optimization surrogate, the quadratic model placed maximum removal at 10% v/v dilution in a flat near-optimal region spanning about 37–40 °C and 74–130 min (maximum 97.7%); a confirmatory experiment at 40 °C and 100 min, inside this region, gave 97.4 ± 0.4% ($n = 5$) against 97.5% predicted. For small replicated adsorption datasets, the held-out unit should match the intended use of the model, and predictions away from the sampled levels require models that extrapolate smoothly.

Keywords: machine learning; model selection; cross-validation; data leakage; Bayesian optimization; multi-objective optimization; adsorption; vinasse; SHAP

Check for updates

Academic Editors: Xiaoyu Jiang, Yuanqiang Zhou, Le Yao and Zheren Zhu

Received: 23 September 2026
Revised: 25 October 2026
Accepted: 2 November 2026
Published: 7 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As the demand for sustainable remediation technologies increases, attention has shifted toward materials that are abundant, renewable, and compatible with circular economy principles. In this context, agricultural residues offer a promising resource base for developing environmentally friendly treatment systems. Within this circular infrastructure, lignocellulosic materials derived from agro-industrial residues have become increasingly relevant as low-cost, high-efficiency bioadsorbents [1–3]. Using dead biomass for contaminant capture by bioadsorption offers a "green" and competitive alternative to conventional adsorbents such as commercial activated carbon, whose production and regeneration are often costly.

Several studies have demonstrated the potential of agro-industrial residues as low-cost biosorbents for dye removal. For example, Kerrou et al. [4] and Andrade Siqueira et al. [5] evaluated sugarcane bagasse for methylene blue removal, while Zein et al. [6] used banana stems to biosorb the same dye.

Other studies have focused on modified materials to improve adsorption performance. For instance, El Bery et al. [7] prepared activated carbons from sugarcane bagasse and sawdust to remove phenol and methylene blue from aqueous solutions. Similarly, Chumpiboon et al. [8] applied treated bagasse fly ash for cationic dye removal from textile wastewater, while Mpatani et al. [9] investigated citrate-modified sugarcane bagasse for methylene blue removal from aqueous solutions. Similarly, Abdulrahman et al. [10] produced activated carbon from sugarcane bagasse and applied it to remove phenol from petroleum refinery wastewater.

Sen [11] provided an updated overview of studies that used raw, modified, and treated agricultural solid-waste biomass as adsorbents for removing heavy metal ions from aqueous media. This review discussed materials such as fruit peels, seeds, husks, barks, straws, and biochars derived from these residues.

The sugarcane industry plays a fundamental socioeconomic role in Latin America, and specifically in Panama, with production exceeding two million tons per harvest season [12]. However, this sector generates large volumes of by-products such as sugarcane bagasse and liquid effluents known as vinasses. The latter pose a critical environmental risk because of their recalcitrant nature and high concentration of phenolic compounds, which are toxic to aquatic life and difficult to biodegrade [2,13]. Despite its potential, there has been little research in Panama examining the valorization of sugarcane bagasse as a biosorbent for treating industrial effluents.

Machine learning (ML) has become a common tool for modeling how biosorbent characteristics affect adsorption capacity [14], predicting removal efficiency, and optimizing operating variables such as adsorbent dose, contact time, temperature, pH, and initial concentration. Reported models range from multiple linear regression and response-surface models to support-vector machines, tree ensembles, Gaussian processes, and artificial neural networks, and several studies have concluded that flexible, high-capacity models—particularly neural networks—provide the most accurate predictions [15,16].

In some reports, however, conclusions are frequently drawn from small experimental datasets evaluated with a single random train/test split. Two issues arise from this methodology. First, experimental adsorption datasets are commonly obtained in replicate: the same nominal condition is measured several times, and the spread among replicates creates an irreducible error that no model can be expected to remove. When such replicated data are randomly partitioned, replicates of the same condition can land in both the training and the test set, so a sufficiently flexible model can memorize condition-specific values and appear far more accurate than it truly is—a form of data leakage. Second, without a reference for the intrinsic measurement noise, a low test error is hard to interpret; it may reflect genuine predictive skill or merely an optimistic split.

A more defensible protocol is to (i) group all replicates of a condition together so that whole conditions, not individual measurements, are held out; (ii) tune and evaluate models with nested cross-validation to avoid optimistic bias from hyperparameter selection; and (iii) judge the resulting error against the replicate (pure-error) noise floor. Pure-error estimation and lack-of-fit testing are long established in response surface methodology for replicated factorial designs [17–19], and data leakage through non-independent splits is a recognized cause of over-optimistic results in ML-based science [20], including chemistry [21]; what ML studies of adsorption often omit is the use of the pure error as the reference for cross-validated error. This is the approach we previously applied to a related enzyme-activity prediction problem [22], and which we extend here to an adsorption dataset. Beyond prediction, a common practical goal is to use the fitted model to identify operating conditions that maximize removal; when the selected model is used as a surrogate, this optimization inherits the model's reliability, making rigorous model selection a prerequisite for trustworthy optimization.

Through this study, we aim to contribute to knowledge on valorizing sugarcane bagasse as an adsorbent through machine learning. We conducted a rigorous model-selection study and surrogate-based optimization to remove phenolic compounds (expressed as tannins) from a sugarcane vinasse sample using citric-acid-modified sugarcane bagasse. We obtained bagasse and vinasse samples from local sugarcane-processing facilities in Panama. The contributions are fourfold: (1) a repeated, nested, replicate-aware cross-validation benchmark of eight regressors spanning shallow to deep families, referenced to the replicate pure error and to a classical lack-of-fit test; (2) a quantitative comparison of three validation designs—naive random splitting, grouping by condition, and holding out entire dilution levels—that separates the effect of replicate leakage from that of extrapolation to unseen levels, with a memorization baseline that shows why leakage helps little here; (3) an interpretation of the fitted model through its coded coefficients and analysis of variance, checked with permutation importance and SHAP, identifying the operating variable that governs removal; and (4) a surrogate-based optimization that uses the selected model to find the conditions maximizing removal and to characterize the relationship between removal and adsorption capacity. Relative to our previous study [22], which introduced grouped nested cross-validation against the pure error, the new elements are contributions (2) and (4) and the link between them: the validation design is chosen to match how the surrogate will be used for optimization, and the resulting optimum is confirmed experimentally. The adsorbent preparation and physical characterization are reported separately [2]; here, the focus is predictive modeling and optimization.

2. Materials and Methods

2.1. Dataset

The dataset corresponds to batch adsorption of phenolic compounds (as tannins) from sugarcane vinasse onto citric-acid-modified sugarcane bagasse, at fixed pH 3 with an adsorbent dose of 0.1 g in 20 mL and 150 rpm in a Shel Lab SI4 incubator shaker (Sheldon Manufacturing Inc., Cornelius, OR, USA). It follows a full-factorial design over three operating variables—temperature (30, 35, and 40 °C), contact time (30, 60, 90, 120, 150, 180, and 210 min), and vinasse dilution (10, 20, 30, 40, and 50% v/v)—with each of the 105 conditions measured in triplicate, giving 315 records. Vinasse dilution is the controlled variable; because undiluted vinasse contained 210.78 mg/L of phenolic compounds, each dilution level corresponds to an initial concentration C_o of 21.1, 42.2, 63.2, 84.3, and 105.4 mg/L, respectively. Throughout this paper, "dilution" denotes this variable and C_o its value in mg/L; both refer to the same factor. The response is the removal of phenolic compounds

($R\%$), calculated with Equation (1). The mean and standard deviation of removal for all 105 conditions are given in Table S1.

$$R\% = \frac{\left(C_o - C_f\right)}{C_o} \times 100 \quad (1)$$

where C_o and C_f (mg/L) are the initial and final concentrations of phenolic compounds in the liquid phase, both determined by the Folin–Denis method (AOAC 952.03) [23] and expressed as tannic acid equivalents. C_o and C_f therefore refer to total phenolic compounds quantified as tannins. We measured absorbance with a BioMate 6 UV-Vis spectrophotometer (Thermo Scientific, Cambridge, UK).

The experiments were not performed on a single day. Sampling at seven contact times (30–210 min), along with filtering each sample, makes it impractical to process all conditions continuously. In addition, equipment constraints required running the three temperatures (30, 35 and 40 °C) in separate sessions. All dilutions for the original replicates and the confirmatory assays were prepared from a single vinasse lot (stock C_0 = 210.78 mg/L) stored at 7 °C, using the same batch of bagasse. Because sessions took place on different days, this design cannot fully separate temperature and day-to-day variability, which we acknowledge as a limitation.

Citric-acid-modified sugarcane bagasse was prepared from bagasse collected in the artisanal panela production sector of Chiriquí, Panama. A previous study has reported the preparation and preliminary characterization of the material [2]. Briefly, we washed the bagasse, air-dried it for five days, ground it, and sieved it through a 500 µm test sieve to obtain uniform 500 µm particles. Raw bagasse was mixed with 10% w/v aqueous citric acid solution, sonicated for 30 min, and then heated at 100 °C for 20 min. We filtered the material, dried it at 45 °C for 48 h, and stored it in a humidity-controlled environment. Characterization included pHpzc and surface acid and basic group determination, FTIR spectra, and SEM micrographs. A local distillery donated vinasse samples.

2.2. Predictive Models

We compared eight regressors spanning a range of model complexity levels: ridge (regularized linear) regression; a second-order (quadratic) polynomial model with ridge regularization; support-vector regression with a radial-basis-function kernel (SVR) [24]; random forest [25]; gradient boosting [26]; and a Gaussian-process regressor with an RBF or Matérn kernel selected in the inner cross-validation [27]. Finally, two multilayer-perceptron neural networks [28] were included, a small one (one or two hidden layers of 8–32 neurons) and a deeper one (two or three hidden layers of 16–128 neurons). Temperature, contact time, and vinasse dilution were the only inputs. Inputs were standardized for all models except the tree ensembles, and the response was also standardized for the neural networks. The neural networks (scikit-learn MLPRegressor) were trained either with the L-BFGS quasi-Newton optimizer [29] (full batch, up to 3000 iterations, convergence tolerance 10^{-4}, no early stopping) or with Adam [30] (mini-batches of min (200, n) records, which at this sample size is effectively the full training set; initial learning rate 10^{-3} or 10^{-2}; up to 3000 epochs; early stopping on a 10% validation split after 50 epochs without improvement). Weights were initialized with the Glorot uniform scheme [31] using a fixed seed. The optimizer, activation function (ReLU or tanh), architecture, and L2 penalty were tuned as hyperparameters. The search grids and the selected values for all models are listed in Table A1 (Appendix A); the values selected in every outer fold and the convergence diagnostics of the neural networks are given in Tables S2 and S3.

2.3. Nested Cross-Validation and the Noise Floor

Because every condition was measured in triplicate, a random partition of the 315 records places replicates of the same condition in both the training and test folds, leaking information [32]. We therefore grouped all resampling by experimental condition [33]: the 105 conditions, not the 315 records, were randomly assigned to folds, so that the three replicates of a condition were always held out together. (Random assignment matters: scikit-learn's GroupKFold without shuffling assigns groups in a fixed order, and with labels that sort by factor level this can make each fold coincide with one level of a factor.) Models were assessed by nested cross-validation [34]: an outer five-fold grouped split provided the performance estimate, and within each outer training set an inner three-fold grouped split was used to tune hyperparameters by grid search (minimum RMSE). Out-of-fold predictions were pooled across the outer folds to compute the root-mean-square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). The procedure was repeated five times with different random assignments of conditions to folds, and we report the mean and standard deviation across repeats. To assess generalization to dilution levels absent from the training data, we also ran the same nested procedure with an outer leave-one-dilution-level-out split (five folds, each holding out all 21 conditions of one dilution level), keeping the inner grouped tuning unchanged. All preprocessing steps (polynomial expansion, standardization of the inputs, and standardization of the response for the neural networks) were part of scikit-learn pipelines refitted within each training fold, so no information from a test fold entered them. Hyperparameters were chosen solely by the inner cross-validation (minimum mean RMSE), and all reported performance values come from outer test folds that were not used for tuning. The choice among the eight model families was then made by comparing these outer estimates; with eight candidates compared on the same folds, this choice carries a small selection optimism, which is negligible here because the leading models differ by less than their uncertainty. To express that uncertainty with respect to the finite set of conditions, we computed 95% cluster-bootstrap confidence intervals (2000 resamples of the 105 conditions) of each model's out-of-fold RMSE and of its paired difference from the quadratic model, using the out-of-fold squared errors averaged over the five repeats.

The replicate (pure-error) noise floor was estimated as the root mean of the within-condition variances:

$$RMSE_{floor} = \sqrt{\frac{1}{G}\sum {s_g}^2} \quad (2)$$

where ${s_g}^2 = \Sigma_i(y_{gi} - \bar{y}_g)^2/(n_g - 1)$ is the sample variance (denominator $n_g - 1$) of the $n_g = 3$ replicate measurements y_{gi} of condition g, the sum in Equation (2) runs over $g = 1, \ldots, G$, and $G = 105$ is the number of conditions. The pure-error RMSE is the error expected from a model that predicts the true mean of every condition. It is therefore a practical reference for the attainable accuracy (referred to below as the noise floor) rather than a strict bound: cross-validated estimates from a finite sample can fall slightly below it by chance. Its uncertainty was expressed by a 95% confidence interval from the χ^2 distribution with 210 degrees of freedom and, because replicate scatter differs between conditions, by the same cluster bootstrap over conditions.

To relate the cross-validated error of the selected model to classical response surface methodology, the quadratic model was also fitted by ordinary least squares on coded variables (each factor scaled to $-1 \ldots +1$), and its residual sum of squares was partitioned into pure error ($315 - 105 = 210$ degrees of freedom) and lack of fit ($105 - 10 = 95$ degrees of freedom); lack of fit was tested with $F = MS_{LOF}/MS_{PE}$ [17,18]. Because a new measurement carries replicate scatter, the largest R^2 attainable for predicting new replicates is approximately $1 - {s_{pe}}^2/{s_y}^2$, where ${s_y}^2$ is the total variance of removal.

2.4. Comparison of Validation Designs

To quantify the effect of the validation design itself, each model was evaluated with fixed hyperparameters (the values most frequently selected in the nested procedure; Table A1) under three five-fold schemes: (i) a naive split, in which the 315 records were shuffled individually, ignoring the replicate structure; (ii) the grouped split described above; and (iii) leave-one-dilution-level-out [32,33]. Schemes (i) and (ii) were repeated with 20 random seeds, and the paired difference $\Delta R^2 = R^2_{\text{naive}} - R^2_{\text{grouped}}$ was summarized by its mean and its 2.5–97.5th percentile range across seeds. Because the repeats share the same data, this range describes the variability due to the random split, not a sampling confidence interval. As a reference for what pure memorization of replicates can achieve, we also evaluated a replicate-memorizing predictor which, under the naive split, predicts each test record as the mean of the replicates of the same condition present in the training folds.

2.5. Model Interpretation

The selected model was interpreted directly through its coefficients on coded variables, with 95% confidence intervals from the ordinary least-squares fit, and through an analysis of variance (ANOVA) by term [17,18]. Permutation importance and SHAP values were computed as model-agnostic checks that apply equally to every model in the comparison. Permutation importance [25] was computed on held-out data: within each outer fold of the grouped cross-validation (five folds × five repeats), the model was fitted to the training conditions, and the increase in test RMSE after randomly permuting one variable in the held-out fold was averaged over 30 permutations. SHAP (SHapley Additive exPlanations) values [35] were computed for the final model, i.e., the selected model refitted to all 315 records, which is the model used as the optimization surrogate. Because the model has only three inputs, we used the exact (interventional) Shapley explainer of the SHAP library with an independent masker whose background set was the 105 distinct design points, and we explained all 315 records; importance was summarized as the mean absolute SHAP value per variable. As a check, SHAP values were also computed for each held-out fold using the model fitted without that fold.

2.6. Optimization

Because the selected surrogate is a quadratic polynomial, its maximum over the experimental box (temperature 30–40 °C, contact time 30–210 min, dilution 10–50% v/v) was obtained directly by bounded quasi-Newton maximization (L-BFGS-B) from 27 starting points and checked on a dense grid (41 × 181 × 81 points). Bayesian optimization is not needed for a closed-form surrogate; we ran it only as a check (gp_minimize of scikit-optimize [36–38]; Gaussian process with a Matérn $\nu = 5/2$ kernel; 10 Latin hypercube initial points [39] followed by 40 expected-improvement iterations [40] with $\xi = 0.01$; noise level 10^{-10}; 10 random seeds), because the same workflow would be required for surrogates without a closed form, such as a Gaussian process or a neural network. To describe the relationship that matters for treatment design, we computed the adsorption capacity $q = (C_o - C_f)V/m = C_o RV/(100m)$ (mg/g), with $V/m = 0.2$ L/g, from the predicted removal R, and extracted from a dense grid (41 × 91 × 161 points) the Pareto front that maximizes both R and q. These q values are derived from predicted removal and the fixed dose; they are not separate measurements.

2.7. Software

All analyses were performed in Python with scikit-learn (models, cross-validation, permutation importance) [41], scikit-optimize (Bayesian optimization), the SHAP library (feature attribution) [35], and NumPy [42], pandas [43], and Matplotlib [44]. Random

seeds were fixed for reproducibility, and the code and data are available in the repository listed in the Data Availability Statement.

We used scikit-learn 1.9.1, scikit-optimize 0.10.2, SHAP 0.53.0, NumPy 2.5.3, pandas 3.0.5, and Matplotlib 3.11.2, under Python 3.13.

3. Results and Discussion

3.1. Dataset Characteristics

The removal rate ranged from 62.96% to 100% (mean 78.84%, standard deviation 9.93%). Vinasse dilution (i.e., initial concentration) was the most influential variable: the Pearson correlation between vinasse dilution and removal was −0.858 (Spearman $\rho = -0.868$), indicating that more concentrated vinasse was removed less efficiently. The gain from higher temperature was greatest at 10% v/v between 30 and 35–40 °C, and smallest at 50% v/v (Figure 1a). Temperature showed a weak positive association with removal ($r = 0.138$; $\rho = 0.144$), and contact time was essentially uncorrelated with it ($r = -0.012$; $\rho = -0.012$). Within each dilution, removal did not change systematically with contact time (Figure 1b), indicating that equilibrium was reached by the first sampling time (30 min) or earlier. Because no samples were taken before, the initial kinetics could not be resolved. At 50% v/v, removal at 30 min was slightly lower than at later times, suggesting a somewhat slower approach at the highest loading.

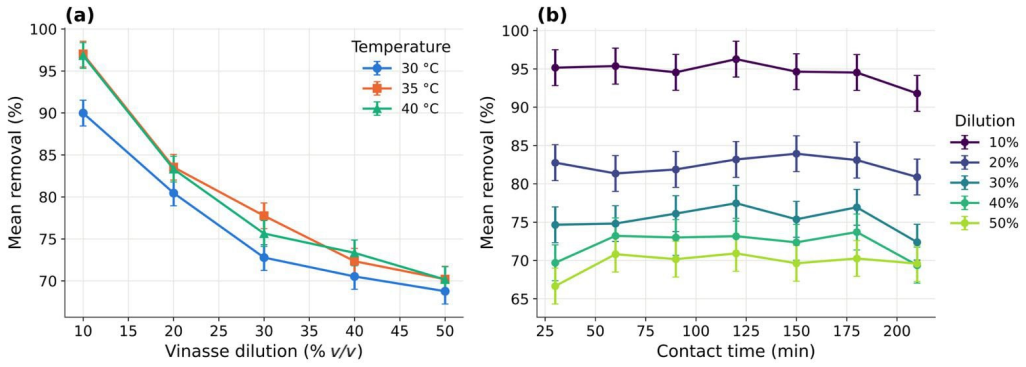


Figure 1. Mean removal (**a**) versus vinasse dilution at each temperature (means over contact time) and (**b**) versus contact time at each dilution (means over temperature). Error bars: 95% confidence intervals of the means based on the pure-error mean square (210 degrees of freedom).

A decrease in mean removal at 210 min was also observed for the more dilute levels (10–30% v/v). Because we ran a blank in parallel at each contact time and temperature and subtracted it from Cf, release from the bagasse is unlikely to explain it; at these low initial concentrations, it may reflect combined variability in the sample, blank, and method. The replicate noise floor was 3.56% removal (95% CI 3.25–3.94% from the χ^2 distribution, 2.80–4.34% from the bootstrap over conditions), i.e., about one third of the overall standard deviation, indicating substantial intrinsic measurement scatter that limits the accuracy that can be expected from any model.

3.2. Model Selection Under Nested Cross-Validation

Table 1 reports the nested cross-validation performance of the eight models, ordered by their grouped RMSE, and Figure 2 compares the errors with the noise floor. With conditions grouped across folds, the non-linear models performed almost identically: gradient boosting, the quadratic model, SVR, the small neural network, the deep neural network, the Gaussian process, and random forest all reached RMSE between 3.89% and 4.10% (R^2 0.83–0.85), i.e., within about 0.5 percentage points of the 3.56% noise floor. The differences among

them (at most 0.21 percentage points) are comparable to their variation across repeats (SD 0.02–0.11), so their ranking is not stable. Cluster-bootstrap intervals support this (Table S4): the RMSE of the quadratic model has a 95% CI of 3.38–4.45%, which overlaps that of the pure error, and the paired differences from the quadratic model include zero for gradient boosting, SVR, the small neural network, and the deep neural network, whereas the Gaussian process and random forest were slightly but consistently worse (by 0.17 and 0.20 percentage points). Only the linear model was clearly worse (RMSE 5.06%, R^2 0.74), confirming that the response has genuine curvature. For predicting new combinations of the sampled levels, model complexity therefore neither helped nor hurt: beyond a smooth quadratic trend, the dataset contains little structure that a flexible model could exploit, and the remaining error is dominated by replicate scatter.

Table 1. Nested cross-validation performance of the eight models (pooled out-of-fold predictions). Grouped: conditions randomly assigned to folds, mean ± SD over five repeats. Unseen level: leave-one-dilution-level-out. Replicate pure error = 3.56% RMSE.

Model	Family	RMSE (%), Grouped	MAE (%), Grouped	R^2, Grouped	RMSE (%), Unseen Level	R^2, Unseen Level
Gradient boosting	tree ensemble	3.89 ± 0.07	2.81 ± 0.09	0.846 ± 0.006	9.23	0.13
Quadratic	polynomial	3.90 ± 0.02	2.89 ± 0.02	0.845 ± 0.001	4.58	0.79
SVR (RBF)	kernel	3.97 ± 0.11	2.86 ± 0.14	0.839 ± 0.009	5.01	0.74
Neural net (small)	neural network	4.00 ± 0.03	2.93 ± 0.03	0.838 ± 0.003	5.81	0.66
Neural net (deep)	neural network	4.01 ± 0.09	2.93 ± 0.09	0.837 ± 0.007	7.16	0.48
Gaussian process	Gaussian process	4.07 ± 0.06	2.99 ± 0.07	0.831 ± 0.005	5.55	0.69
Random forest	tree ensemble	4.10 ± 0.08	2.99 ± 0.09	0.829 ± 0.007	9.51	0.08
Ridge (linear)	linear	5.06 ± 0.04	4.07 ± 0.04	0.739 ± 0.004	7.06	0.49

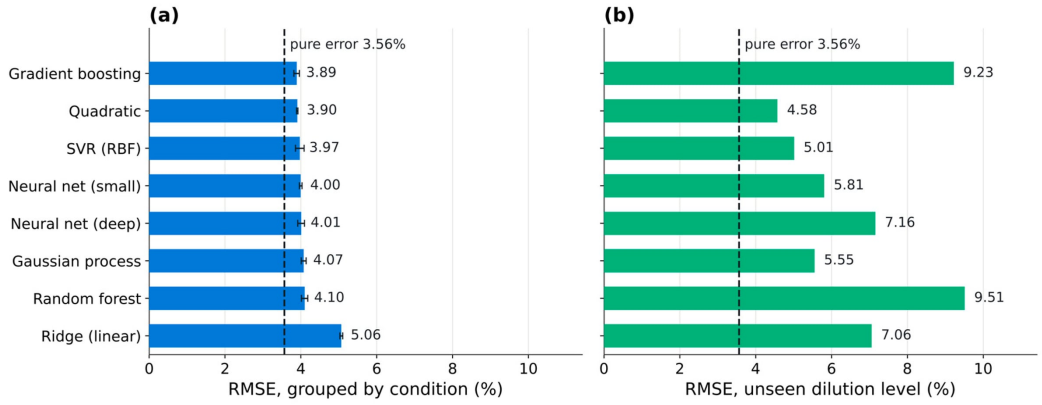


Figure 2. Nested cross-validation RMSE of the eight models (**a**) with conditions grouped across folds (mean ± SD over five repeats) and (**b**) with one dilution level held out per fold. Dashed line: replicate pure error (3.56%).

The picture changed when whole dilution levels were held out (right-hand columns of Table 1; Figure 2b). The quadratic model was then the most accurate (RMSE 4.58%, R^2 0.79), followed by SVR, the Gaussian process, and the small neural network, whereas random forest and gradient boosting failed (R^2 0.08 and 0.13). The tree ensembles had their largest errors when the 10% and 20% levels were held out (RMSE 12–13%), where removal changes most steeply with dilution (Figure 1a): piecewise-constant models cannot predict values outside the range of their training targets and can only interpolate stepwise between the remaining levels. The deeper neural network was less reliable in this setting (R^2 0.48), mainly because it performed poorly when the 10% level was held out (RMSE

13.7%); its results also depended strongly on the hyperparameters selected in each fold (R^2 0.77 with fixed hyperparameters; Table 2). Every other model was significantly worse than the quadratic model in this setting (paired differences 0.44–4.94 percentage points, all 95% CIs excluding zero; Table S4).

Table 2. Cross-validated R^2 under three validation designs at fixed hyperparameters (Table A1). Naive and grouped: mean over 20 random seeds; ΔR^2: mean with 2.5–97.5th percentile range over seeds. Unseen level: leave-one-dilution-level-out; values differ from Table 1 because hyperparameters are fixed here rather than re-tuned within each fold.

Model	R^2, Naive	R^2, Grouped	ΔR^2 (Naive − Grouped)	R^2, Unseen Level
Gradient boosting	0.86	0.85	0.008 [0.000, 0.021]	0.11
Quadratic	0.85	0.84	0.005 [0.001, 0.013]	0.79
SVR (RBF)	0.85	0.84	0.009 [0.002, 0.017]	0.78
Neural net (small)	0.85	0.84	0.012 [0.003, 0.022]	0.74
Neural net (deep)	0.86	0.85	0.010 [−0.003, 0.024]	0.77
Gaussian process	0.85	0.83	0.013 [0.005, 0.025]	0.72
Random forest	0.84	0.83	0.008 [0.000, 0.013]	0.10
Ridge (linear)	0.75	0.74	0.009 [0.001, 0.019]	0.49

3.3. The Decisive Role of the Validation Protocol

Table 2 and Figure 3 compare the three validation designs at fixed hyperparameters. The naive random split inflated R^2 only marginally relative to grouping by condition: the mean ΔR^2 ranged from +0.005 to +0.013 across models, with 2.5–97.5th percentile ranges spanning −0.003 to +0.025. The reason is quantitative. Under the naive split, 96% of the test records had at least one replicate of the same condition in the training folds, so leakage was pervasive; however, a predictor that simply memorized those replicates reached an RMSE of 4.96%. Even in its most favorable case, with two replicates in the training folds, such a predictor has an expected error of $\sigma_{pe}\sqrt{(1 + 1/2)} = 4.36\%$, which is higher than the error of a smooth model fitted to neighboring conditions under grouped validation (quadratic RMSE 3.90%). When replicate scatter is large relative to the variation between neighboring conditions, memorizing replicates offers no advantage, and the naive estimate is only slightly optimistic. Leakage would matter more for datasets with more precise replicates or a rougher response surface, and its size can be checked as we have done here.

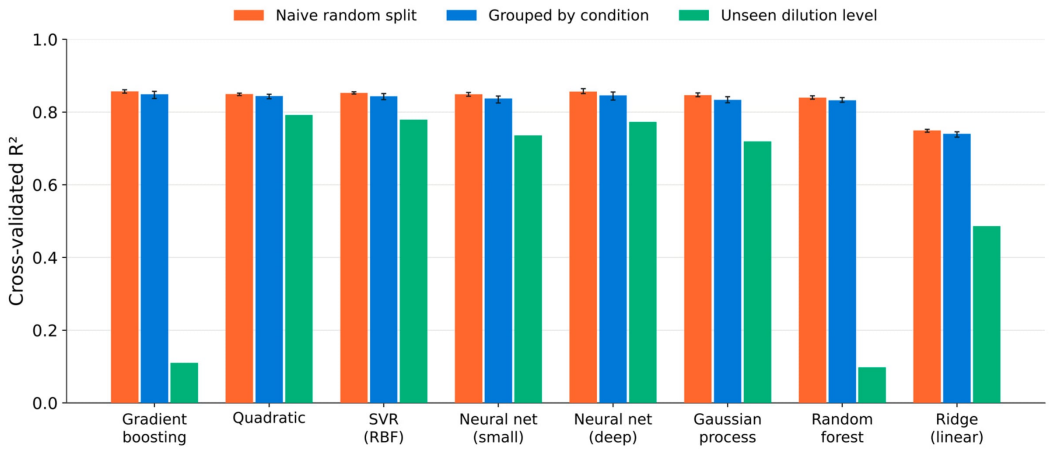


Figure 3. Cross-validated R^2 of each model (fixed hyperparameters) under a naive random split and a split grouped by condition (mean and 2.5–97.5th percentile range over 20 seeds), and with one dilution level held out per fold.

By contrast, the choice of held-out unit had a large effect. Holding out entire dilution levels reduced R^2 by 0.05 for the quadratic model but by about 0.7 for the tree ensembles (Table 2). This is not a leakage effect but a different prediction task: the model must predict a level of the dominant variable that it has never seen. This task is relevant here, because a surrogate used for optimization is queried across the whole operating window, and the optimum lies at the edge of the sampled range (Section 3.6). Reports that rank complex models as the best adsorption predictors should therefore be read in light of the validation design: random splits mainly test interpolation between replicated conditions, which most flexible models do well, and say little about predictions away from the sampled levels. The validation design should match the intended use of the model—grouping by condition for predicting new combinations within the sampled levels, and holding out whole levels when the model will be used between or beyond them. We therefore use "leakage" only for the first effect—information shared between training and test records of the same condition—and treat prediction at a truly new condition, whether a new combination of sampled levels or an unseen level, as a question of generalization.

3.4. Performance of the Selected Model

Figure 4a shows the quadratic model's out-of-fold predictions from the first repeat of grouped cross-validation against the measured values. The apparent flattening—high measurements under-predicted at 10% *v*/*v* and low ones over-predicted at 50% *v*/*v*—is mostly the expected effect of replicate scatter on such a plot: because each measurement contains noise that no model can predict, even a model that predicts every condition mean exactly would give a slope of predicted on measured values of about $1 - {s_{pe}}^2/{s_y}^2 \approx 0.87$, close to the observed slope of 0.85. Residuals grouped by dilution (Figure 4b) show only small systematic offsets (mean −0.8 to +0.9 percentage points across the five levels), and residuals were uncorrelated with fitted values ($r = -0.03$).

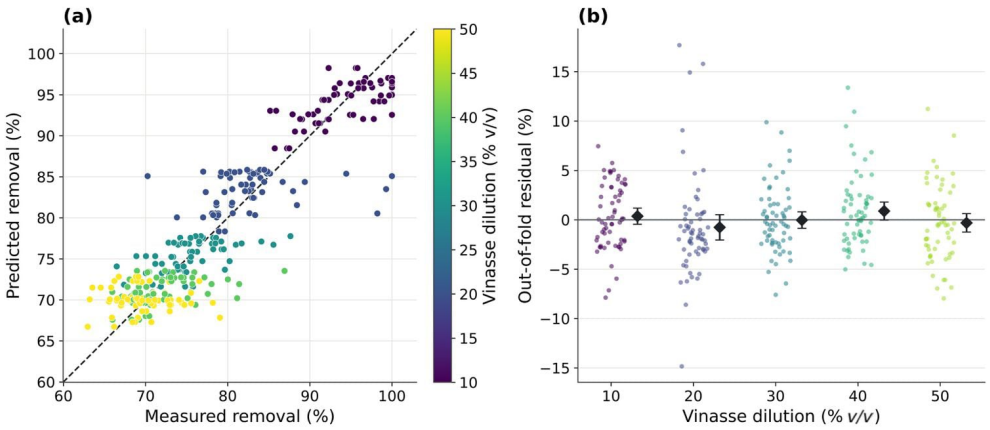


Figure 4. (**a**) Out-of-fold predicted versus measured removal for the quadratic model (grouped cross-validation, first repeat; dashed line: identity; color: vinasse dilution). (**b**) Out-of-fold residuals by dilution level; diamonds: mean ± 95% confidence interval.

The classical analysis of the same model (Table 3) agrees. Lack of fit is statistically significant but small ($F_{95,210} = 1.42$, $p = 0.02$): of the residual mean square of 14.3, about 12.7 is pure error, so the lack-of-fit component has a standard deviation of roughly 1.3 percentage points, which adds in quadrature to the pure error ($\sqrt{(3.56^2 + 1.3^2)} \approx 3.78\%$, the residual standard deviation). Treating dilution as a five-level factor removed part of it ($F = 3.94$, $p = 0.02$), indicating that the dilution effect is not exactly quadratic, but significant lack of fit remained ($p = 0.04$). The in-sample R^2 of the quadratic model was 0.86 and its grouped cross-validated R^2 0.845, against an attainable value of about 0.87 for new replicates. The

model therefore captures most, but not all, of the systematic variation, and the remaining lack of fit is smaller than the replicate scatter.

Table 3. Analysis of variance of the quadratic model fitted by ordinary least squares on coded variables. Term rows give the extra sum of squares when the term (or group of terms) is removed from the full model. T: temperature; t: contact time; D: vinasse dilution; SS: sum of squares; MS: mean square.

Source	df	SS	MS	*F*	*p*
Linear	3	23,396.3	7798.8	544.77	<0.001
Quadratic	3	3010.9	1003.6	70.11	<0.001
Interaction	3	170.8	56.9	3.98	0.008
D (dilution)	1	22,802.0	22,802.0	1592.79	<0.001
D^2	1	2455.8	2455.8	171.55	<0.001
T (temperature)	1	589.8	589.8	41.20	<0.001
T^2	1	276.3	276.3	19.30	<0.001
t (time)	1	4.6	4.6	0.32	0.573
t^2	1	278.7	278.7	19.47	<0.001
T × D	1	126.3	126.3	8.82	0.003
T × t	1	0.0	0.0	0.00	0.965
t × D	1	44.5	44.5	3.11	0.079
Residual	305	4366.3	14.3		
Lack of fit	95	1705.5	18.0	1.42	0.020
Pure error	210	2660.8	12.7		
Total (corrected)	314	30,944.3			

Because removal cannot exceed 100% (12 of the 315 measurements were exactly 100%), we also fitted the quadratic model to logit-transformed removal. Under grouped cross-validation, it was less accurate than the untransformed model for every clipping threshold tried for the 100% values (RMSE 4.34–5.81% for thresholds of 5% to 0.5%, against 3.90%), and with the 0.5% threshold its optimum moved to the edge of the design (40 °C, 30 min). Within the design region, the untransformed model never predicts more than 97.7%, so it never reaches the bound, and we retained it.

The learning curve (Figure 5) was obtained, within each of the 25 outer folds, by fitting the model to random subsets of whole training conditions (so that replicates stayed together) and evaluating it on the held-out conditions. The held-out error fell monotonically from 4.82% with 51 training records to 3.87% with 252, approaching the noise floor, so additional conditions within the same design would yield only marginal gains.

3.5. Interpretation of the Selected Model

Table 4 gives the fitted coefficients on coded variables, so each linear coefficient is the change in removal from the center to the edge of the factor's range. Dilution has by far the largest effect (linear coefficient −12.0 percentage points, 95% CI −12.6 to −11.4) with significant curvature (+6.7), so removal falls steeply from 10% to 30% v/v and then levels off. Temperature has a smaller positive effect with a maximum inside the range (linear +1.7, quadratic −2.0), contact time acts only through curvature (quadratic −2.4; linear term not significant, $p = 0.57$), and the only significant interaction is temperature × dilution (−1.1), meaning that the benefit of a higher temperature is largest at the most dilute levels (Figure 1a). Permutation importance on held-out conditions and SHAP values (Figure 6) rank the variables in the same order (dilution: RMSE increase 9.07%, mean |SHAP| 7.89%; temperature: 0.60%, 1.54%; contact time: 0.19%, 0.83%), and SHAP values computed on held-out folds gave the same ranking. For a quadratic model on an orthogonal design, these model-agnostic summaries add little to the coefficients; we report them because they apply unchanged to the non-parametric models compared in Section 3.2.

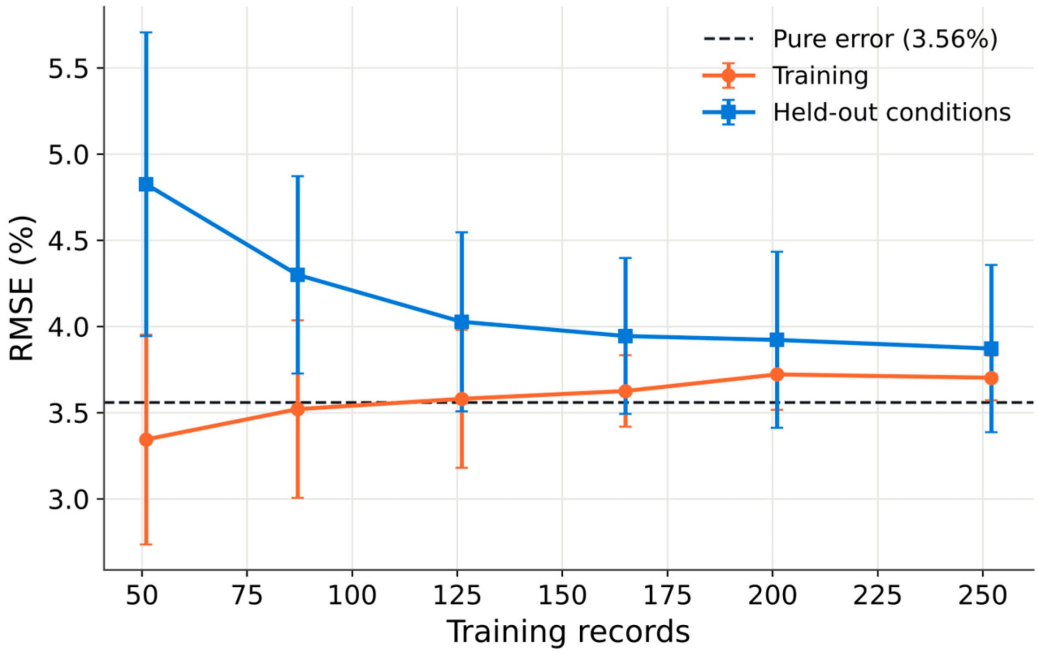


Figure 5. Learning curve of the quadratic model: training and held-out RMSE when the model is fitted to random subsets of the training conditions (mean ± SD over five folds × five repeats). Dashed line: replicate pure error.

Table 4. Coefficients of the quadratic model on coded variables (each factor scaled to −1 … +1; T: temperature, t: contact time, D: vinasse dilution), with 95% confidence intervals from the ordinary least-squares fit (305 degrees of freedom).

Term	Coefficient (%)	95% CI (%)	*p*
Intercept	77.92	76.91 to 78.92	<0.001
T	1.68	1.16 to 2.19	<0.001
t	−0.18	−0.81 to 0.45	0.573
D	−12.03	−12.63 to −11.44	<0.001
T^2	−1.99	−2.88 to −1.10	<0.001
t^2	−2.44	−3.53 to −1.35	<0.001
D^2	6.67	5.67 to 7.68	<0.001
T × t	−0.02	−0.79 to 0.75	0.965
T × D	−1.10	−1.82 to −0.37	0.003
t × D	0.80	−0.09 to 1.69	0.079

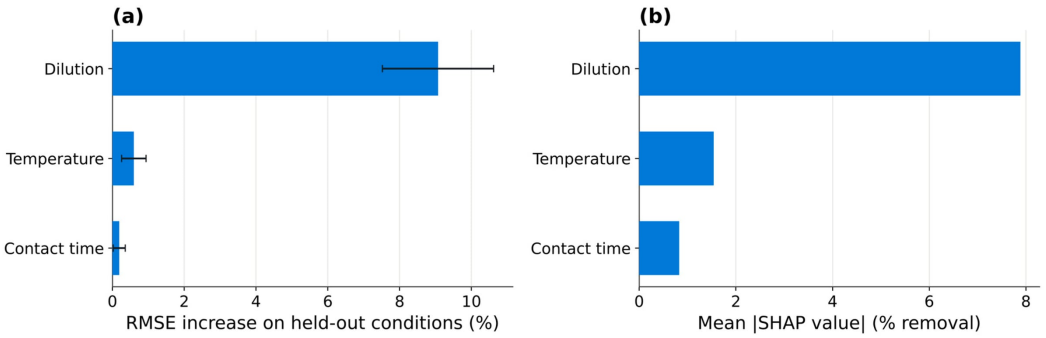


Figure 6. Variable importance for the quadratic model: (**a**) permutation importance on held-out conditions (mean ± SD over 25 folds × 30 permutations); (**b**) mean absolute SHAP value of the final model over all 315 records (exact explainer). Both rank dilution first, temperature second, and contact time last.

Two features of the design matter when interpreting these results. First, the design is full factorial, so the three inputs are exactly uncorrelated; the importance of dilution is not confounded with temperature or time, and the coefficients in Table 4 can be read independently. Second, removal is the ratio of the amount adsorbed to the amount initially present. With a fixed adsorbent dose, a higher initial concentration lowers the fraction removed even as the amount adsorbed per gram rises: measured adsorption capacity increases with dilution almost monotonically (Spearman $\rho = 0.979$), opposite to the correlation between removal and dilution ($\rho = -0.868$). The dominant importance of dilution therefore largely reflects the loading of a fixed mass of adsorbent and the way the response is defined, rather than evidence of a specific mechanism; dilution also changes the vinasse matrix (e.g., color, salts, competing organics), which this design cannot separate from concentration. For this reason, Section 3.6 considers removal and capacity together.

3.6. Optimization of Operating Conditions

The maximum of the quadratic surrogate over the experimental box lies at the lower bound of dilution (10% v/v, $C_o \approx 21$ mg/L), 38.6 °C, and 102 min, with a predicted removal of 97.7%; the dense-grid evaluation and all 10 Bayesian-optimization runs reached the same point (38.4–38.8 °C, 101–104 min, 97.7%). The optimum is flat (Figure 7a): predicted removal stays within 0.25 percentage points of the maximum over a roughly elliptical region spanning about 37–40 °C and 74–130 min, a difference far smaller than the replicate scatter, so any condition in this region is effectively optimal. Dilution is driven to its lower bound because it dominates removal. This boundary optimum illustrates why the leave-one-level-out results matter: the 10% and 20% levels are where the tree ensembles had their largest errors, so a surrogate chosen on interpolation accuracy alone could have misplaced the optimum. The reliability of surrogate-based optimization depends on the surrogate's accuracy in the region where the optimum lies, which in turn depends on how well the training samples cover that region [45].

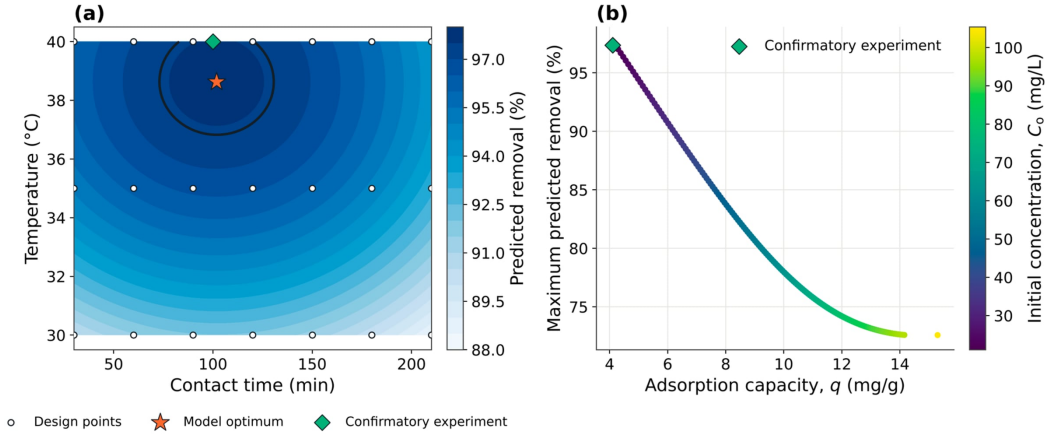


Figure 7. (**a**) Predicted removal of the quadratic model at 10% v/v dilution as a function of contact time and temperature; black contour: 0.25 percentage points below the maximum (star); white circles: design points; diamond: confirmatory experiment. (**b**) Pareto front of maximum predicted removal versus adsorption capacity q, colored by initial concentration; diamond: confirmatory experiment (measured removal). The gap near $q = 15$ mg/g arises because the quadratic model predicts a shallow minimum of removal near 48% v/v, so dilutions of about 46–50% v/v are dominated.

To assess how reliably the optimum is located, we refitted the model to 500 bootstrap resamples of the 105 conditions and located the maximum of each refitted surface on a grid. Dilution was at its lower bound in every resample; the optimal temperature ranged from 37 to 40 °C and the optimal contact time from 72 to 146 min (95% intervals), with a maximum

predicted removal of 96.4–99.1%. In 97.5% of resamples, the predicted removal at the tested condition (40 °C, 100 min) was within 1.0 percentage point of the resample's maximum. The temperature and time of the optimum are therefore uncertain within the flat region, whereas the removal attainable in that region is consistently high.

Maximizing removal alone simply selects the most diluted feed, which also uses the adsorbent least effectively. Figure 7b therefore shows the relationship between removal and adsorption capacity: along the Pareto front, q rises from 4.1 mg/g at Co ≈ 21 mg/L (97.7% removal) to 15.3 mg/g at Co ≈ 105 mg/L (72.6% removal), with optimal temperatures of 36–39 °C and contact times of 102–132 min. Within the range studied, treating less diluted vinasse thus removes almost four times more phenolic compounds per gram of adsorbent, while removal decreases by about 25 percentage points.

3.7. Experimental Validation of the Optimum

Conditions in the optimal region (40 °C, 100 min, and 10% v/v dilution) were tested experimentally in a fresh set of assays ($n = 5$); the final surrogate predicts 97.5% there, 0.14 percentage points below its maximum. The measured removal, 97.4 ± 0.4% (mean ± SD), agreed closely with the prediction (difference −0.2 percentage points) and lay between the means measured in the original design at 90 min (94.6 ± 2.0%) and 120 min (98.8 ± 1.7%) at the same temperature and dilution. The mean of the five replicates lay within the 95% confidence interval of the predicted mean response (96.1–98.9%), and all five replicates lay within the 95% prediction interval for a single new observation (89.9–105.1%; Figure 8). Both intervals were obtained from the ordinary least-squares fit of the same quadratic model (residual standard deviation 3.78%, 305 degrees of freedom); the small ridge penalty of the final model has a negligible effect on the prediction, and the upper limit of the prediction interval exceeds 100% because the model does not constrain removal to its physical range. Because 100 min lies between the sampled 90 and 120 min, this confirmation is an interpolation within the design, not an out-of-design test.

The confirmatory replicates were much more consistent (SD 0.4%) than the original design replicates, whose pure-error SD was 2.3% at 10% v/v dilution and 3.56% overall (χ^2 test of the variance ratio, $p = 0.002$). Replicate scatter in the original design was also uneven across dilution levels (pure-error SD 2.3–5.5%). The original replicates therefore include sources of variation that were smaller in the confirmatory assays, and the intervals in Figure 8, which use the original design's residual standard deviation (3.78%), are conservative for these replicates.

3.8. Implications and Limitations

The results support two conclusions of different scopes. The first is specific to this dataset: with three operating variables, a smooth response, and replicate scatter of 3.56% removal, a quadratic model is as accurate as any of the more flexible models for predicting new combinations of the sampled levels, and it is the most accurate when predicting unseen dilution levels; it is therefore the appropriate surrogate here. We do not claim that simple models generally outperform complex ones—with more variables, wider ranges, stronger non-linear effects, or larger datasets, flexible models may well be superior. The second conclusion is methodological and more general: the performance estimate depends on what is held out. Replicates must be kept together, but in this dataset that precaution changed R^2 by only about 0.01 (mean ΔR^2 0.005–0.013), whereas holding out whole levels of the dominant variable changed the ranking of the models substantially. Comparisons of ML models for adsorption should therefore state the held-out unit, report the replicate pure error, and, when the model is used for optimization, test it in the region where it will be queried.

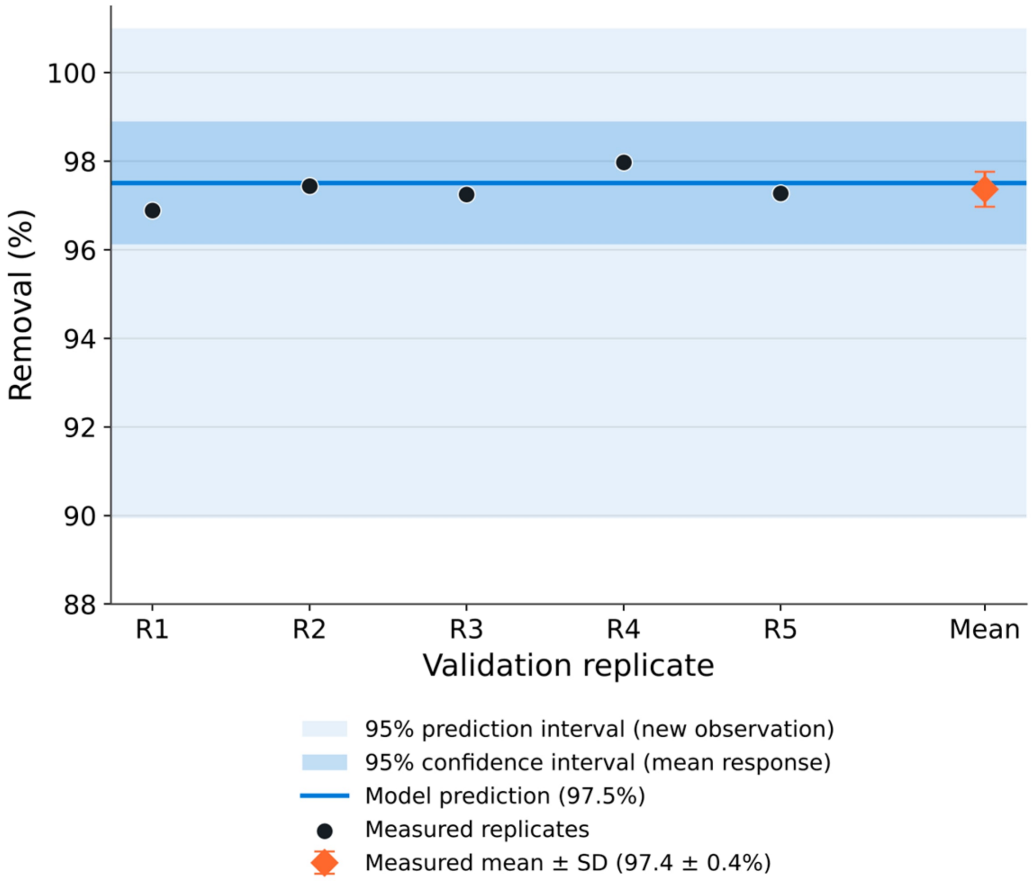


Figure 8. Confirmatory experiment in the optimal region (40 °C, 100 min, 10% v/v dilution). Points: five measured replicates; diamond: their mean ± SD. Line: prediction of the quadratic model; dark band: 95% confidence interval of the mean response; light band: 95% prediction interval for a single new observation (both from the ordinary least-squares fit of the quadratic model; the prediction interval is drawn up to 101%).

This study has limitations. It concerns a single adsorbent–adsorbate system with three operating variables over relatively narrow ranges, at fixed pH and adsorbent dose, and contact time had little effect within the sampled range of 30–210 min; datasets with stronger kinetic effects (systems far from equilibrium) or with pH as an additional variable would have a more complex response surface and might favor more flexible models. The quadratic model shows a small but significant lack of fit, mainly in the shape of the dilution effect, and it does not constrain removal to 100%; neither affects the optimal region here, but a more flexible dilution term or a bounded-response model could be preferable for other datasets. The leave-one-level-out test probes only the dominant variable. The optimum was confirmed experimentally at a single condition inside the design (Section 3.7); a broader validation across the operating window, in particular at higher initial concentrations, would further strengthen the recommendations.

4. Conclusions

We conducted a replicate-aware model-selection study and surrogate-based optimization to predict phenolic-compound removal from sugarcane vinasse by citric-acid-modified sugarcane bagasse. For this dataset, repeated nested cross-validation with conditions grouped across folds showed that all non-linear models, from a quadratic model to neural networks, reached similar accuracy (RMSE 3.89–4.10% against a replicate pure error of 3.56%), while a naive random split overstated R^2 by only about 0.01, because replicate mem-

orization offers no advantage when replicate scatter is large. Holding out whole dilution levels, by contrast, separated the models sharply: the quadratic model remained the most accurate (RMSE 4.58%), while the tree ensembles failed. A classical lack-of-fit test showed a small but significant departure from the quadratic form ($F_{95,210} = 1.42$, $p = 0.02$), smaller than the replicate scatter. Vinasse dilution (initial concentration) was the dominant variable. The maximum of the quadratic surrogate lies at 10% v/v dilution in a flat near-optimal region spanning about 37–40 °C and 74–130 min, and a confirmatory experiment in this region (40 °C, 100 min) agreed with the prediction (97.4 ± 0.4% measured vs. 97.5% predicted). Along the Pareto front, adsorption capacity rises from 4.1 to 15.3 mg/g as removal falls from 97.7% to 72.6%. Beyond this dataset, we recommend that ML studies of adsorption keep replicates together during validation, report the replicate pure error and a lack-of-fit test as references, and choose the held-out unit to match the intended use of the model, especially when the model will be used to optimize conditions near or beyond the edges of the sampled range. Whether simple or flexible models are preferable must be determined for each dataset rather than assumed.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/pr14213605/s1, Table S1: Experimental conditions with the mean and standard deviation of removal ($n = 3$); Table S2: Hyperparameters selected in each outer fold of the nested cross-validation; Table S3: Convergence diagnostics of the neural networks; Table S4: Cluster-bootstrap confidence intervals of out-of-fold RMSE and of paired differences from the quadratic model.

Author Contributions: Methodology, software, formal analysis, visualization, writing, N.A.J. and J.C.M.; investigation/data curation, N.A.J., M.M. and M.C.; writing—review and editing, N.A.J., M.M., M.C., J.C.M. and R.E.C.; conceptualization and supervision, R.E.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT), Panama. Grant APY-NI-2024A-14.

Data Availability Statement: The dataset (315 records) and the Python code used for all analyses are available at https://github.com/juanky2797/vinasse-phenol-ml (accessed on 14 October 2026). The mean and standard deviation of removal for the 105 experimental conditions are given in Table S1.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ANOVA	Analysis of variance
AOAC	Association of Official Analytical Chemists
ARD	Automatic relevance determination
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
MAE	Mean absolute error
ML	Machine learning
RBF	Radial basis function
RMSE	Root-mean-square error
SD	Standard deviation
SHAP	SHapley Additive exPlanations
SVR	Support-vector regression
UV–Vis	Ultraviolet–visible spectroscopy

Appendix A

Table A1. Hyperparameter grids searched in the inner cross-validation and the most frequently selected values across the 25 outer folds (five folds × five repeats) of the grouped nested cross-validation; where two settings were selected equally often, both are shown, and the first was used for Table 2 and the final model. The neural networks used up to 3000 iterations (L-BFGS) or epochs (Adam; batch size min (200, n), early stopping with patience 50 on a 10% validation split), Glorot-uniform initialization with a fixed seed, and standardized inputs and response.

Model	Grid Searched in the Inner Cross-Validation	Most Frequent Selection (Count/25)
Gradient boosting	trees ∈ {100, 300}; learning rate ∈ {0.03, 0.1}; max depth ∈ {1, 2, 3}; subsample ∈ {0.7, 1.0}	trees = 100; learning rate = 0.1; depth = 2; subsample = 0.7 or trees = 300; learning rate = 0.03; depth = 2; subsample = 1.0 (6/25 each)
Quadratic	$\alpha \in \{0.001, 0.01, 0.1, 1, 10\}$	$\alpha = 0.01$ or $\alpha = 0.001$ (10/25 each)
SVR (RBF)	$C \in \{1, 10, 100, 1000, 10{,}000\}$; $\gamma \in \{0.01, 0.1, 1\}$; $\varepsilon \in \{0.1, 1, 2\}$	$C = 10{,}000$; $\gamma = 0.01$; $\varepsilon = 2$ (6/25)
Neural net (small)	hidden layers ∈ {(16), (32), (16, 8)}; L-BFGS: activation ∈ {ReLU, tanh}, L2 ∈ {0.001, 0.01, 0.1, 1, 10}; Adam: ReLU, L2 ∈ {0.001, 0.1}, learning rate ∈ {0.001, 0.01}, early stopping	(16, 8); L-BFGS; tanh; L2 = 1.0 or (32); L-BFGS; ReLU; L2 = 1.0 (6/25 each)
Neural net (deep)	hidden layers ∈ {(64, 32), (64, 32, 16), (128, 64, 32)}; other settings as for the small network	(64, 32); L-BFGS; ReLU; L2 = 1.0 (6/25)
Gaussian process	kernel ∈ {RBF, Matérn $\nu = 3/2$, Matérn $\nu = 5/2$}, each ARD × constant + white noise; kernel hyperparameters by marginal likelihood (2 restarts)	RBF (15/25)
Random forest	300 trees; max depth ∈ {2, 3, 6, none}; min. samples per leaf ∈ {1, 3, 9}; max. features ∈ {1, 2, 3}	max depth = 3; leaf = 3; features = 3 or max depth = none; leaf = 3; features = 2 (6/25 each)
Ridge (linear)	$\alpha \in \{0.01, 0.1, 1, 10, 100\}$	$\alpha = 1$ (11/25)

References

1. Sarker, T.C.; Azam, S.M.G.G.; El-Gawad, A.M.A.; Gaglione, S.A.; Bonanomi, G. Sugarcane Bagasse: A Potential Low-Cost Biosorbent for the Removal of Hazardous Materials. *Clean Technol. Environ. Policy* **2017**, *19*, 2343–2362. [CrossRef]
2. Jaramillo-González, N.A.; Miranda, M.; Ceballos, J.; Caballero, R.E. Efficient Removal of Tannins from Vinasse with Citric Acid-Modified Sugarcane Bagasse: Process Optimization via Response Surface Methodology, Kinetic and Equilibrium Studies. *Appl. Environ. Res.* **2027**, *49*, 1–21. [CrossRef]
3. Mpatani, F.M.; Han, R.; Aryee, A.A.; Kani, A.N.; Li, Z.; Qu, L. Adsorption Performance of Modified Agricultural Waste Materials for Removal of Emerging Micro-Contaminant Bisphenol A: A Comprehensive Review. *Sci. Total Environ.* **2021**, *780*, 146629. [CrossRef]
4. Kerrou, M.; Bouslamti, N.; Raada, A.; Elanssari, A.; Mrani, D.; Slimani, M.S. The Use of Sugarcane Bagasse to Remove Organic Dyes from Wastewater. *Int. J. Anal. Chem.* **2021**, *2021*, 5570806. [CrossRef]
5. Andrade Siqueira, T.C.; da Silva, I.Z.; Rubio, A.J.; Bergamasco, R.; Gasparotto, F.; Paccola, E.A.d.S.; Yamaguchi, N.U. Sugarcane Bagasse as an Efficient Biosorbent for Methylene Blue Removal: Kinetics, Isotherms and Thermodynamics. *Int. J. Environ. Res. Public Health* **2020**, *17*, 526. [CrossRef]
6. Zein, R.; Akmal, C.; Safni, S.; Fauzia, S.; Ramadhani, P. Banana Stem (*Musa balbisiana* Colla) as Potential Biosorbent to Remove Methylene Blue Dye in Wastewater: Isotherm, Kinetic, Thermodynamic Studies and Its Application. *Appl. Environ. Res.* **2023**, *45*, 15. [CrossRef]
7. El-Bery, H.M.; Saleh, M.; El-Gendy, R.A.; Saleh, M.R.; Thabet, S.M. High Adsorption Capacity of Phenol and Methylene Blue Using Activated Carbon Derived from Lignocellulosic Agriculture Wastes. *Sci. Rep.* **2022**, *12*, 5499. [CrossRef]

8. Chumpiboon, A.; Thongsubsai, K.; Pongsiri, T.; Knijnenburg, J.T.N.; Ngernyen, Y. Removal of Cationic Dye from Textile Wastewater Using Treated Bagasse Fly Ash: An Industrial Waste. *Eng. Appl. Sci. Res.* **2022**, *49*, 381–394. [CrossRef]
9. Mpatani, F.M.; Aryee, A.A.; Kani, A.N.; Wen, K.; Dovi, E.; Qu, L.; Li, Z.; Han, R. Removal of Methylene Blue from Aqueous Medium by Citrate Modified Bagasse: Kinetic, Equilibrium and Thermodynamic Study. *Bioresour. Technol. Rep.* **2020**, *11*, 100463. [CrossRef]
10. Abdulrahman, M.S.; Alsarayreh, A.A.; Barno, S.K.A.; Abd Elkawi, M.A.; Abbas, A.S. Activated Carbon from Sugarcane as an Efficient Adsorbent for Phenol from Petroleum Refinery Wastewater: Equilibrium, Kinetic, and Thermodynamic Study. *Open Eng.* **2023**, *13*, 10. [CrossRef]
11. Sen, T.K. Agricultural Solid Wastes Based Adsorbent Materials in the Remediation of Heavy Metal Ions from Water and Wastewater by Adsorption: A Review. *Molecules* **2023**, *28*, 5575. [CrossRef]
12. Solís, R. MIDA—Proyección de la Producción Agrícola en Panamá Año Agrícola 2022–2023—Conjunto de Datos. Available online: https://www.datosabiertos.gob.pa/dataset/mida-proyeccion-de-la-produccion-agricola-en-panama (accessed on 22 June 2023).
13. Montiel Rosales, A.; Montalvo Romero, N.; García Santamaría, L.E.; Sandoval Herazo, L.C.; Bautista Santos, H.; Fernández Lambert, G. Post-Industrial Use of Sugarcane Ethanol Vinasse: A Systematic Review. *Sustainability* **2022**, *14*, 11635. [CrossRef]
14. Yang, Y.Y.; Li, Z.L.; Wang, G.; Zhao, X.P.; Crowley, D.E.; Zhao, Y.H. Computational Identification and Analysis of the Key Biosorbent Characteristics for the Biosorption Process of Reactive Black 5 onto Fungal Biomass. *PLoS ONE* **2012**, *7*, e33551. [CrossRef]
15. Hassan, R.; Kazemi, M.R. Machine Learning Frameworks to Accurately Estimate the Adsorption of Organic Materials onto Resin and Biochar. *Sci. Rep.* **2025**, *15*, 15157. [CrossRef]
16. El Mallahi, M.; Akichouh, E.; Elyoussfi, A.; Salhi, A.; Ahari, M.; Amhamdi, H.; Ghalit, M.; Mourabit, F. Artificial Intelligence in Adsorption Process Optimization: A 2014–2024 Bibliometric Analysis of Research Trends and Developments. *Results Eng.* **2025**, *28*, 107777. [CrossRef]
17. Montgomery, D.C. *Design and Analysis of Experiments*, 9th ed.; Wiley: Hoboken, NJ, USA, 2017.
18. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 4th ed.; Wiley: Hoboken, NJ, USA, 2016.
19. Bezerra, M.A.; Santelli, R.E.; Oliveira, E.P.; Villar, L.S.; Escaleira, L.A. Response Surface Methodology (RSM) as a Tool for Optimization in Analytical Chemistry. *Talanta* **2008**, *76*, 965–977. [CrossRef]
20. Kapoor, S.; Narayanan, A. Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns* **2023**, *4*, 100804. [CrossRef]
21. Sheridan, R.P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783–790. [CrossRef]
22. Montenegro, J.C.; Miranda, M.; Torres, J.; Caballero, R.E. Model Selection for Fungal Laccase Activity: Shallow Learners versus Neural Networks under Nested Cross-Validation. *Rev. Acad. Colomb. Cienc. Exactas Fis. Nat.* **2026**, *50*, 132–142. [CrossRef]
23. AOAC International. *Official Methods of Analysis of AOAC International*, 18th ed.; Method 952.03; AOAC International: Gaithersburg, MD, USA, 2005.
24. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
25. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
27. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006. [CrossRef]
28. Akinci, T.C.; Topsakal, O.; Akbas, M.I. Machine Learning Methods from Shallow Learning to Deep Learning. In *Shallow Learning vs. Deep Learning*; Ertuğrul, Ö.F., Guerrero, J.M., Yilmaz, M., Eds.; Springer: Cham, Switzerland, 2024; pp. 1–28.
29. Liu, D.C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45*, 503–528. [CrossRef]
30. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
31. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), Sardinia, Italy, 13–15 May 2010; pp. 249–256.
32. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 15. [CrossRef]
33. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
34. Cawley, G.C.; Talbot, N.L.C. On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.

35. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
36. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; De Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
37. Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian Optimization of Machine Learning Algorithms. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 2951–2959.
38. Head, T.; Kumar, M.; Nahrstaedt, H.; Louppe, G.; Shcherbatyi, I. Scikit-Optimize: Sequential Model-Based Optimization in Python, Version 0.10.2. Available online: https://scikit-optimize.github.io (accessed on 14 October 2026).
39. McKay, M.D.; Beckman, R.J.; Conover, W.J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979**, *21*, 239–245. [CrossRef]
40. Jones, D.R.; Schonlau, M.; Welch, W.J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Glob. Optim.* **1998**, *13*, 455–492. [CrossRef]
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Vanderplas, J.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
42. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array Programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef]
43. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; pp. 56–61. [CrossRef]
44. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
45. HaiChao, C.; Wenlong, H.; Zuyuan, L.; Baiwei, F.; Qiang, Z. Enhancing Surrogate Model Accuracy in Ship Design Optimization through Intelligent Constraint-Aware Sample Selection. *Eng. Appl. Artif. Intell.* **2026**, *163*, 112716. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.