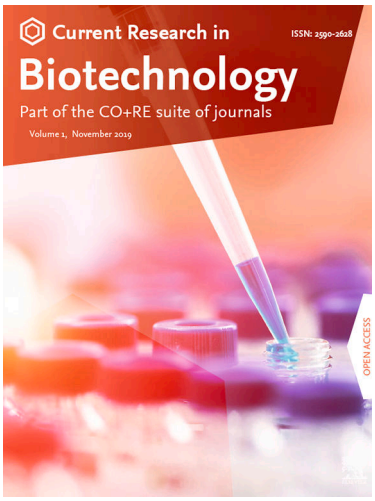
Journal Pre-proofs



Stochastic optimization of Gradient Boosting Decision Trees for interpretable prediction of heavy metal adsorption onto biochar

Mahran Al-Zyoud, Salama A. Mostafa, Ibrahim Khersan, J. Gowrishankar, Prabhat Kumar Sahu, Siya Singla, Sardor Sabirov, Islom. Khudayberganov, Samim Sherzod

PII: S2590-2628(26)00029-8
DOI: https://doi.org/10.1016/j.crbiot.2026.100393
Reference: CRBIOT 100393

To appear in: *Current Research in Biotechnology*

Received Date: 6 March 2026
Revised Date: 1 May 2026
Accepted Date: 25 May 2026

Please cite this article as: M. Al-Zyoud, S.A. Mostafa, I. Khersan, J. Gowrishankar, P.K. Sahu, S. Singla, S. Sabirov, Islom. Khudayberganov, S. Sherzod, Stochastic optimization of Gradient Boosting Decision Trees for interpretable prediction of heavy metal adsorption onto biochar, *Current Research in Biotechnology* (2026), doi: https://doi.org/10.1016/j.crbiot.2026.100393

Stochastic Optimization of Gradient Boosting Decision Trees for Interpretable Prediction of Heavy Metal Adsorption onto Biochar

Stochastic Optimization of Gradient Boosting Decision Trees for Interpretable Prediction of Heavy Metal Adsorption onto Biochar

Mahran Al-Zyoud [1], Salama A. Mostafa [2], Ibrahim Khersan [3], Gowrishankar J [4], Prabhat Kumar Sahu [5], Siya Singla [6], Sardor Sabirov [7], Islom. Khudayberganov [8], Samim Sherzod [9]*

1 Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan

2 Department of Artificial Intelligence, College of Engineering Technology, Alnoor University, Mosul, 41012, Nineveh, Iraq

3 Department of computers Techniques engineering, College of technical engineering, The Islamic University, Najaf, Iraq

4 Department of Computer Science Engineering, School of Engineering and Technology, JAIN (Deemed to be University), Bangalore, Karnataka, India

5 Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha-751030, India

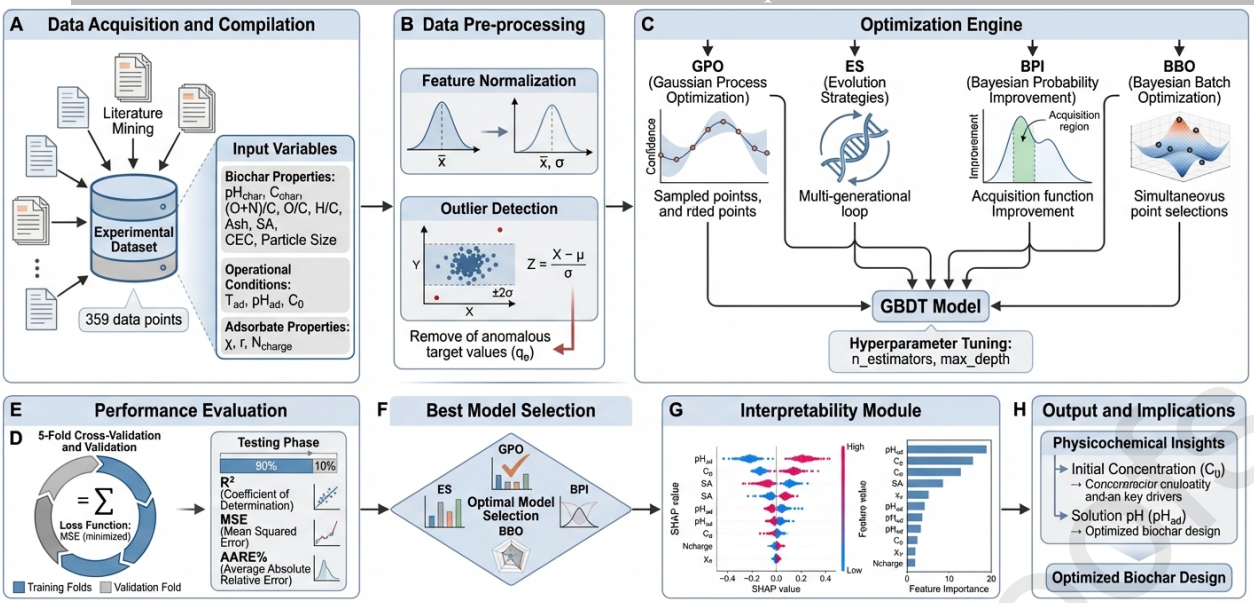
6 Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India

7 Department of General Professional Sciences, Mamun University, Khiva, Uzbekistan

8 Department of chemistry, Urgench state university, Uzbekistan

9 Faculty of Engineering, Nangarhar University, Nangarhar, Afghanistan

***Corresponding Author: samimsherzod@gmail.com**



Highlights

- ✓ This study aimed to develop a rigorous, interpretable machine learning framework for predicting equilibrium adsorption capacity by optimizing Gradient Boosting Decision Trees
- ✓ A comprehensive dataset of 359 experimental points encompassing diverse biochar physicochemical properties and operational conditions was utilized to train models tuned by four distinct heuristics
- ✓ Results demonstrated that while all algorithms achieved near-perfect training fits, the Gaussian Process Optimization framework emerged as the superior architecture, delivering the highest generalization stability with a testing coefficient of determination of 0.9784 and minimized mean squared error
- ✓ initial metal concentration and solution pH function as the dominant thermodynamic drivers, significantly outweighing physical surface morphology in determining adsorption efficacy

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Credit author statement

All authors contributed to this paper.

Mahran Al-Zyoud [1], Salama A. Mostafa [2], Ibrahim Khersan [3], Gowrishankar J [4], Prabhat Kumar Sahu [5], Siya Singla [6], Sardor Sabirov [7], Islom. Khudayberganov [8], Samim Sherzod [9]*

1 Department of Networks and Cybersecurity, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan

2 Department of Artificial Intelligence, College of Engineering Technology, Alnoor University, Mosul, 41012, Nineveh, Iraq

3 Department of computers Techniques engineering, College of technical engineering, The Islamic University, Najaf, Iraq

4 Department of Computer Science Engineering, School of Engineering and Technology, JAIN (Deemed to be University), Bangalore, Karnataka, India

5 Department of Computer Science and Information Technology, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha-751030, India

6 Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India

7 Department of General Professional Sciences, Mamun University, Khiva, Uzbekistan

8 Department of chemistry, Urgench state university, Uzbekistan

9 Faculty of Engineering, Nangarhar University, Nangarhar, Afghanistan

***Corresponding Author: samimsherzod@gmail.com**

Abstract

Predicting the heavy metal adsorption capacity of biochar is a significant challenge due to complex physicochemical mechanisms and the limitations of traditional experimental approaches. This study aimed to develop and validate a robust, interpretable machine learning framework by optimizing Gradient Boosting Decision Trees (GBDT) for this predictive task. Using a comprehensive dataset of 359 experimental points, we compared four hyperparameter optimization heuristics and found that Gaussian Process Optimization (GPO) yielded a model with superior generalization performance. The final GBDT-GPO model achieved a coefficient of determination (R^2) of 0.9784 and a mean squared error (MSE) of 0.0035 on an unseen test set, in contrast to other methods like Evolution Strategies, which showed significant overfitting. Furthermore, Shapley Additive Explanations (SHAP) analysis identified initial metal concentration and solution pH as the dominant factors governing adsorption, outweighing physical properties like surface area. This research establishes a highly accurate and interpretable computational strategy that can guide the rational design of biochar and optimize its application in water treatment.

Keywords: Biochar Adsorption, Gradient Boosting Decision Tree, Bayesian Optimization, Machine Learning Interpretability, Heavy Metal Remediation

1. Introduction

The rapid escalation of industrialization and urbanization over the last century has precipitated a critical global water crisis, characterized prominently by the accumulation of heavy metal pollutants in aquatic ecosystems [1, 2]. Toxic elements such as lead (Pb^{2+}), cadmium (Cd^{2+}), copper (Cu^{2+}), and zinc (Zn^{2+}) are routinely discharged via metallurgical processes, battery manufacturing, and agricultural runoff. Unlike organic contaminants, which can often be degraded biologically, heavy metals are non-biodegradable, persistent, and prone to bioaccumulation within the food web, posing severe neurotoxic and carcinogenic risks to human health even at trace concentrations [3, 4]. Consequently, the development of efficient, scalable, and economically viable remediation technologies has become a paramount objective in environmental engineering. While traditional methods such as chemical precipitation, membrane filtration, and ion exchange have been deployed extensively, they are frequently hampered by high operational costs, incomplete removal at low concentrations, and the generation of secondary toxic sludge that requires further hazardous waste management [5, 6].

In response to these limitations, adsorption has emerged as a superior alternative due to its simplicity, high efficiency, and potential for adsorbent regeneration. Among the plethora of available adsorbents, biochar, a carbon-rich solid product resulting from the thermochemical pyrolysis of biomass under oxygen-limited conditions, has garnered significant scientific attention. Biochar represents a sustainable, carbon-negative solution that valorizes agricultural waste while providing a highly porous structure with a vast surface area [7, 8]. More critically, the surface of biochar is decorated with diverse oxygen-containing functional groups (carboxyl, hydroxyl, and phenolic groups) that facilitate complexation and electrostatic attraction with metal cations. However, the adsorption performance of biochar is not uniform; it is highly dependent on a complex matrix of pyrolysis conditions (temperature, residence time) and feedstock types, which in turn dictate the material's physicochemical properties such as pH, carbon content, and cation exchange capacity (CEC) [9, 10].

The physicochemical complexity of the biochar-heavy metal interface presents a substantial challenge for material design and process optimization. The adsorption capacity (q_e) is governed by highly non-linear interactions between the adsorbent's surface characteristics and the operational environmental conditions, including solution pH, initial metal concentration (C_0), and temperature. Traditional experimental approaches rely heavily on trial-and-error methodologies to optimize these parameters [11, 12]. These conventional methods are not only resource-intensive and time-consuming but often fail to capture the synergistic or antagonistic interactions between multiple variables. For instance, increasing pyrolysis temperature may increase surface area but simultaneously destroy the volatile functional groups necessary for chemisorption, creating a trade-off that is difficult to model using simple linear regression or classical isotherm models (Langmuir or Freundlich) alone [1, 13].

To overcome the limitations of experimental heuristics, Machine Learning (ML) has recently been integrated into environmental science as a powerful tool for predictive modeling. Unlike deterministic models, ML algorithms can learn intricate, non-linear patterns from historical datasets, enabling the accurate prediction of adsorption capacities based on multidimensional input features. Techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forest (RF) have shown promise in mapping biochar synthesis parameters to remediation performance. Among these, Gradient Boosting Decision Trees (GBDT) have emerged as a particularly robust architecture for tabular datasets [14-16]. GBDT operates as an ensemble method, sequentially constructing weak learners (decision trees) where each new tree corrects the residual errors of its predecessors. This additive strategy allows GBDT to achieve superior predictive accuracy and generalization capability compared to single-model approaches, particularly in datasets with high variance and complex feature interactions [17, 18].

However, the efficacy of GBDT models is intrinsically tied to the optimal configuration of their hyperparameters, such as the number of estimators, learning rate, tree depth, and subsampling ratios [19, 20]. In appropriate

hyperparameter settings can lead to severe overfitting, where the model memorizes the training noise, or underfitting, where it fails to capture the underlying chemical trends. The majority of existing literature in this domain relies on rudimentary tuning methods like Grid Search or Random Search [21, 22]. Grid Search is computationally exhaustive and often infeasible for high-dimensional spaces, while Random Search lacks a strategic guidance mechanism, potentially missing global optima. There is a distinct paucity of research applying advanced, stochastic optimization algorithms to fine-tune GBDT models specifically for biochar adsorption [23, 24]. Algorithms such as Bayesian Optimization and Evolutionary Strategies offer intelligent, data-efficient pathways to navigate the hyperparameter landscape, yet their comparative efficacy in environmental modeling remains largely unexplored [25, 26].

Furthermore, a critical gap exists regarding the interpretability of black-box ML models in adsorption science. While high predictive accuracy is desirable, it is insufficient for engineering applications if the physical rationale remains opaque [27, 28]. Environmental chemists require insights into *why* a model predicts a high adsorption capacity to validate that the model is learning thermodynamic principles rather than statistical coincidences. Previous studies have often neglected this aspect, providing metrics without mechanistic explanations. The integration of interpretability frameworks, such as Shapley Additive exPlanations (SHAP), is essential to bridge the divide between computational accuracy and physicochemical reality. SHAP analysis allows for the quantification of the marginal contribution of each input feature, such as surface area versus CEC, thereby confirming whether the model aligns with established mass transfer and surface complexation theories [29, 30].

This study develops a rigorous and integrated computational framework that couples Gradient Boosting Decision Trees (GBDT) with four advanced hyperparameter optimization strategies—Gaussian Process Optimization (GPO), Bayesian Probability Improvement (BPI), Bayesian Batch Optimization (BBO), and Evolution Strategies (ES)—to address persistent limitations in predictive accuracy and interpretability for heavy metal adsorption modeling. Using a comprehensive dataset of 359 experimental observations, the framework systematically evaluates the ability of uncertainty-aware Bayesian methods versus evolutionary heuristics to navigate the highly nonlinear and noisy hyperparameter landscape typical of environmental datasets, with the objective of minimizing predictive error while preserving generalization. Beyond algorithmic comparison, the novelty of this work lies in the integration of Monte Carlo–based outlier detection with SHAP-driven interpretability analysis, ensuring that the resulting models are both statistically robust and physicochemically meaningful. The complete workflow, spanning data acquisition, rigorous preprocessing, optimized model training and validation, and post hoc feature attribution, yields a reliable and interpretable predictive tool for virtual screening and rational design of high-performance biochar adsorbents.

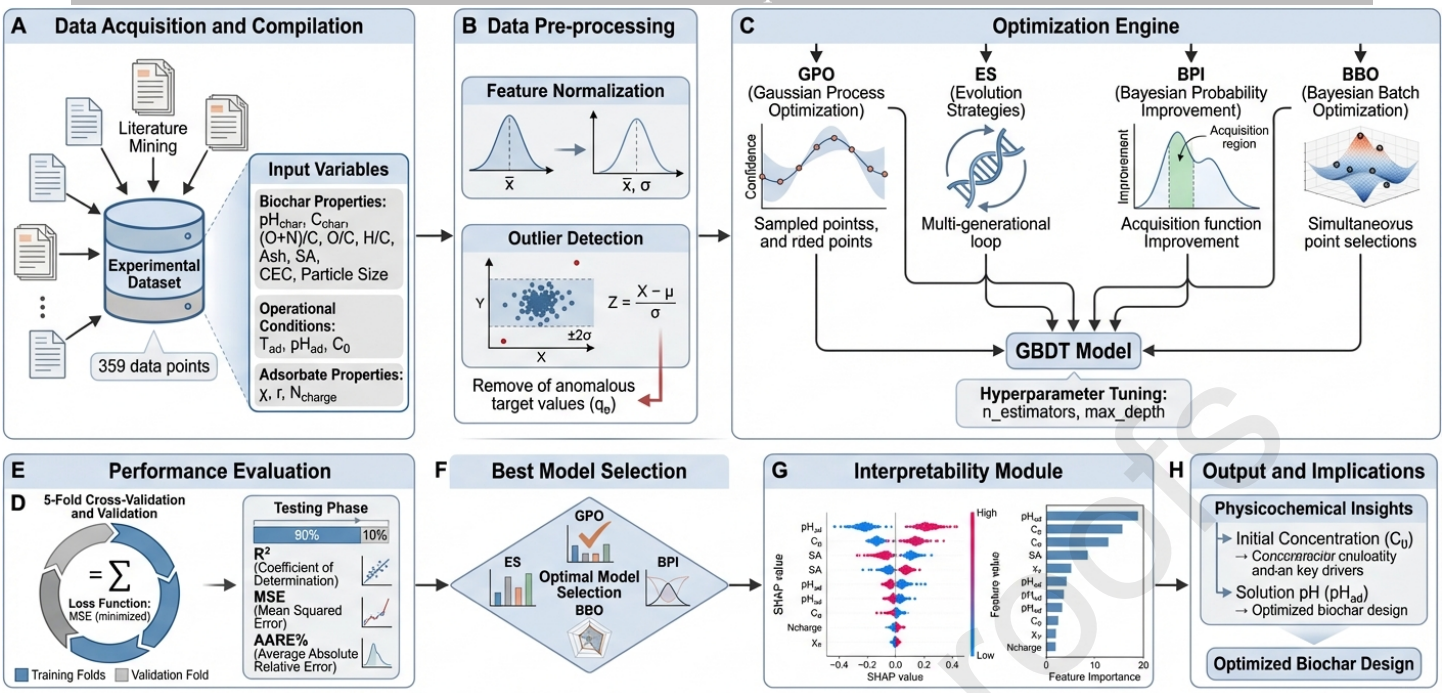


Figure 1. Comprehensive computational workflow employed in this study, illustrating the sequence from data compilation and pre-processing through the stochastic hyperparameter optimization of the GBDT model, followed by rigorous statistical validation and SHAP-based physicochemical interpretability analysis

2. Methodology

Here, the computational frameworks are initially introduced, encompassing the core predictive model and the accompanying optimization algorithms. Subsequently, the compiled dataset comprising 359 datapoints extracted from peer-reviewed journal papers is detailed. This dataset incorporates comprehensive input parameters, specifically pH_char, C_char(wt%), (O+N)/C, O/C, H/C, Ash_char(wt%), SA(m^2/g), CEC(c molc/kg), Particle size (nm), T_ad(°C), pH_ad, C0(mmol/g), χ, r, and Ncharge, which are utilized to forecast the metal adsorption capacity on biochar (q_e (mmol/g)).

2.1 Machine Learning Algorithms

In this study, Gradient Boosting Decision Trees (GBDT) were employed as the core predictive framework. GBDT is a robust ensemble machine learning technique that constructs a predictive model by sequentially combining multiple weak learners, specifically decision trees, to iteratively minimize a specified loss function. This additive approach is highly effective for capturing complex, non-linear interactions inherent in environmental datasets, such as the diverse physicochemical properties of biochar. Because the predictive performance of GBDT is highly sensitive to its internal hyperparameters, such as the number of estimators, learning rate, and maximum tree depth, rigorous tuning is essential to prevent overfitting and ensure optimal generalization.

To efficiently navigate the complex hyperparameter space of the GBDT model, we applied and compared four advanced optimization heuristics. These algorithms represent a mix of probabilistic surrogate-based methods that balance search space exploration with exploitation (GPO, BPI, BBO), and stochastic population-based approaches inspired by biological adaptation (ES). For detailed mathematical formulations, algorithmic control parameters, and the foundational theory governing the GBDT model and these specific optimization strategies, readers are referred to Section S1 in the Supplementary Materials.

2.2 Dataset Description

A comprehensive compilation of 359 distinct experimental datapoints is utilized to establish the predictive framework [14, 21, 22, 31-35]. The intrinsic physicochemical characteristics of the synthesized biochars are quantified through several key input parameters. Specifically, the biochar pH (pH_char, dimensionless), carbon content (C_char, wt%), and ash content (Ash_char, wt%) are included. The significance of pH_char is attributed to its profound influence on the surface charge of the adsorbent and the subsequent electrostatic interactions with target metal ions. Furthermore, C_char dictates the fundamental carbonaceous matrix and structural stability, while Ash_char signifies the presence of inorganic mineral components, which frequently provide auxiliary active sites for precipitation and complexation mechanisms.

To further capture the morphological and chemical complexity of the adsorbents, specific surface area (SA, m^2/g), cation exchange capacity (CEC, c mol_c/kg), and Particle size (nm) are incorporated. The available physical space and internal pore structure for metal entrapment are directly represented by SA and Particle size, whereas CEC serves as a critical indicator of the biochar's capacity to exchange inherent basic cations with heavy metal pollutants. Additionally, elemental molar ratios, namely (O+N)/C, O/C, and H/C (dimensionless), are evaluated to gauge the density of oxygen- and nitrogen-containing surface functional groups, the degree of surface polarity, and the aromaticity of the carbon network, all of which are pivotal for chemisorption and surface complexation interactions.

The environmental and operational conditions governing the adsorption process are delineated by the ambient temperature (T_ad, °C), the pH of the aqueous solution (pH_ad, dimensionless), and the initial metal concentration (C0, mmol/g). The thermodynamics and kinetic mobility of the adsorbate species are highly dependent on T_ad. Simultaneously, pH_ad is recognized as a master variable, dictating both the speciation or hydrolysis state of the metal ions in solution and the protonation state of the biochar functional groups. The initial concentration, C0, is introduced to account for the driving force of the mass transfer process, which significantly impacts the concentration gradient between the aqueous phase and the solid adsorbent interface.

The intrinsic chemical properties of the target metal adsorbates are parameterized through electronegativity (χ), ionic radius (r), and ionic charge (Ncharge). These factors are critical for predicting adsorption affinities based on the principles of hard-soft acid-base theory, spatial steric hindrances, and the magnitude of electrostatic attraction. Ultimately, the singular target output variable is the equilibrium metal adsorption capacity on the biochar (q_e, mmol/g). The accurate determination and modeling of q_e are of paramount environmental importance, as this metric quantitatively defines the maximum efficacy and practical viability of a specific biochar material for remediating heavy metal-contaminated aquatic systems.

The multidimensional distribution and pairwise interrelationships of the aforementioned input and output variables are comprehensively visualized in Figure 2. The diagonal elements of the generated scatter matrix display kernel density estimations, which are utilized to illustrate the frequency distributions and variance of individual parameters across the compiled dataset. Concurrently, the off-diagonal bivariate scatter plots are presented to elucidate the highly non-linear, multifaceted correlations existing among the independent variables and the target q_e. The high degree of dispersion and the distinct lack of simple linear dependencies observed in this matrix strongly substantiate the necessity of deploying advanced, non-parametric machine learning algorithms to accurately map this intricate variable space.

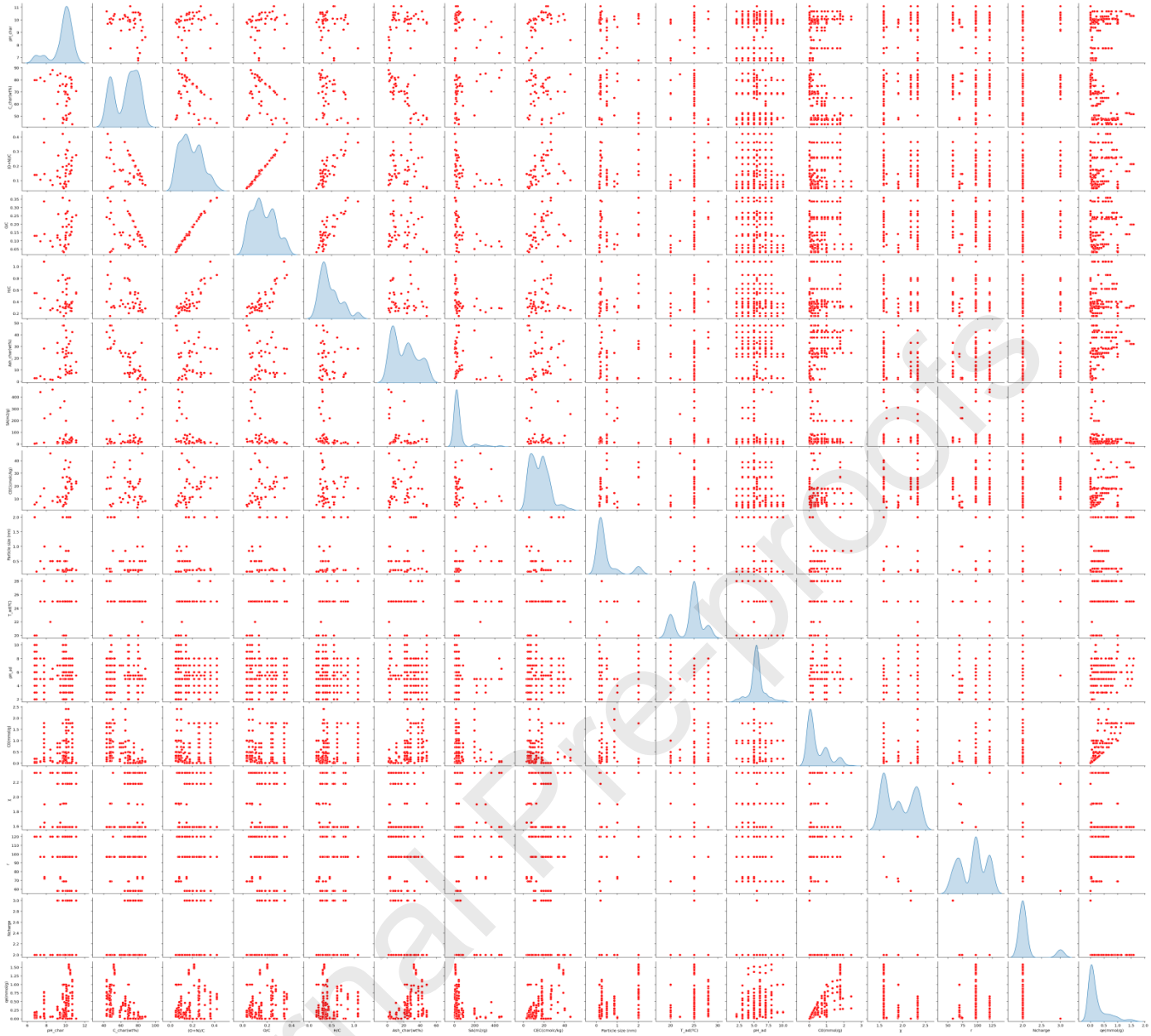


Figure 2. Scatter matrix visualization of the dataset, illustrating univariate kernel density estimations on the diagonal and pairwise bivariate interrelationships for all considered input and output variables.

2.3 Evaluation methods

For the integrity of the model evaluation and strictly prevent data leakage, the dataset was separated in 90% training subsets and 10% unseen test subsets. It must be emphasized that this 10% test set was completely isolated and was not involved in any capacity during the hyperparameter tuning process. To optimize GBDT models, a 5-fold cross-validation (CV) framework was applied exclusively within the 90% training data. Rather than relying solely on exhaustive grid search, advanced optimization algorithms (GPO, ES, BPI, and BBO) were deployed to efficiently navigate the hyperparameter space. These algorithms utilize the average validation error derived from the 5-fold CV on the training subset as their objective function. Once the optimal hyperparameter configuration was identified through this isolated cross-validation process, the final model was trained on the full 90% training set and subsequently evaluated on the untouched 10% test set to provide an unbiased assessment of its predictive capability [36, 37].

Performance for optimized algorithms are quantitatively evaluated through 3 indicators. The R^2 metric is utilized to evaluate the proportion of the variance in the dependent variable that is predictable from the independent variables, serving as a primary indicator of the model's goodness-of-fit. The mathematical formulation for R^2 is defined as follows, where $y_i$represents the actual experimental values, $\hat{y}_i$ denotes the predicted values generated by the model, $\grave{y}$signifies the mean of the observed data, and Nindicates the total number of data points evaluated.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \grave{y})^2}$$

In addition to R^2, the absolute and relative deviations of the predictions are measured via MSE and AARE%, respectively. The MSE calculates the average of the squares of the errors, thereby heavily penalizing larger discrepancies between the predicted and actual values to ensure structural stability in the model. Conversely, AARE% provides a percentage-based estimation of the average relative deviation, offering an intuitive understanding of the model's predictive error across varying scales of the target variable. The equations for these two-error metrics are presented below, utilizing the previously defined parameters y_i, $\hat{y}_i$, and N.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\text{AARE\%} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Upon the completion of the training and testing phases, the computed statistical metrics are reported systematically to provide a transparent evaluation of the model's performance. To accurately detect potential overfitting or underfitting phenomena, the R^2, MSE, and AARE% values are documented separately for the training and testing datasets. Furthermore, a total performance metric is calculated and presented for each statistical measure. This comprehensive total is derived strictly as a weighted average of the training and testing metrics, proportioned precisely according to the 90% and 10% data split, thereby reflecting the overall accuracy and generalization capability of the developed predictive models across the entirety of the compiled dataset [38, 39].

2.4 Optimizing the hyperparameters

To raise the performance and capability of GBDTs, adequate optimization phases are performed. Structure of the constructed models are fundamentally ruled by several critical hyperparameters that must be meticulously calibrated. Specifically, the optimization procedure encompasses the adjustment of the total number of sequential boosting stages (n_estimators), the maximum allowable depth for each individual regression tree (max_depth), and the shrinkage parameter (learning_rate), which mathematically dictates the step size utilized at each iteration to minimize the residual loss. Furthermore, to enhance the model's structural robustness against overfitting and to introduce vital stochastic variance into the training process, spatial and sample constraints are simultaneously tuned; these include the fraction of observations used to fit base learner, the subset proportion of features considered when determining optimum nodes split, the minimum requisite samples to authorize min_samples_split, and the minimum discrete samples dictated to form min_samples_leaf.

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}\left(\mathcal{M}_\lambda(\mathcal{D}_{train}^{(k)}), \mathcal{D}_{val}^{(k)}\right)$$

In the presented objective function, λ^* designates the ultimately selected, optimal hyperparameter configuration extracted from the bounded multidimensional search space Λ. Furthermore, Krepresents the total number of cross-validation folds, $\mathcal{L}$denotes the specified loss function evaluated over the iterative predictions, and $\mathcal{M}_\lambda$represents the designated GBDT model trained on the k-th training fold $\mathcal{D}_{\text{train}}^{(k)}$ and subsequently validated against the corresponding isolated validation fold $\mathcal{D}_{\text{val}}^{(k)}$. To effectively navigate this complex, high-dimensional parameter space and locate λ^*, advanced optimization heuristics. Instead of relying on computationally exhaustive or purely random searches, these algorithms strategically sample specific configurations from Λ, evaluate the cross-validated objective function, and dynamically continuously adjust their sampling trajectories based on historical performance distributions, thereby methodically isolating the most effective GBDT architecture for the compiled dataset.

2.5 Target Outlier Identification

Statistical integrities of compiled datasets are rigorously evaluated prior to model training to identify anomalous observations within the target variable space that could potentially skew the predictive surface. To achieve this, probabilistic boundaries are established, and the distribution of the equilibrium adsorption capacity (q_e) is visualized, as presented in Figure 3. The frequency histogram (Figure 3A) demonstrates a distinctly right-skewed distribution, indicating that the majority of the biochar materials exhibit low to moderate adsorption capacities, while a minority demonstrate exceptionally high performance. To mathematically quantify these extremes, a standard score (Z-score) thresholding methodology is typically employed alongside robust visualization techniques, where deviations are measured relative to the central tendency. The standard deviation boundaries, frequently derived or validated through iterative Monte Carlo simulations, are utilized to delineate the expected operational variance from statistical anomalies [40, 41].

$$Z_i = \frac{y_i - \mu}{\sigma}$$

Within this formulation, Z_i represents the standard score for a given observation, y_i is the specific experimental target value (q_e), μ signifies the arithmetic mean of the target variable across the dataset, and σ denotes the calculated standard deviation. The complementary boxplot (Figure 3B) further elucidates the presence of these extreme values, represented by the dense cluster of points positioned significantly above the upper quartile whisker. Scientifically, in the context of biochar adsorption phenomena, these identified mathematical outliers frequently do not correspond to experimental errors; rather, they represent genuine, albeit extreme, physical conditions [42, 43]. For instance, these elevated q_e values are predominantly generated under conditions of exceptionally high initial adsorbate concentrations, which maximize the thermodynamic driving force, or through the utilization of highly engineered, ultra-porous biochars. The identification of these points is critical, as tree-based algorithms like GBDT, while inherently robust to target outliers compared to parametric models, must still be meticulously regularized during the tuning phase to prevent overfitting to these extreme minority phenomena.

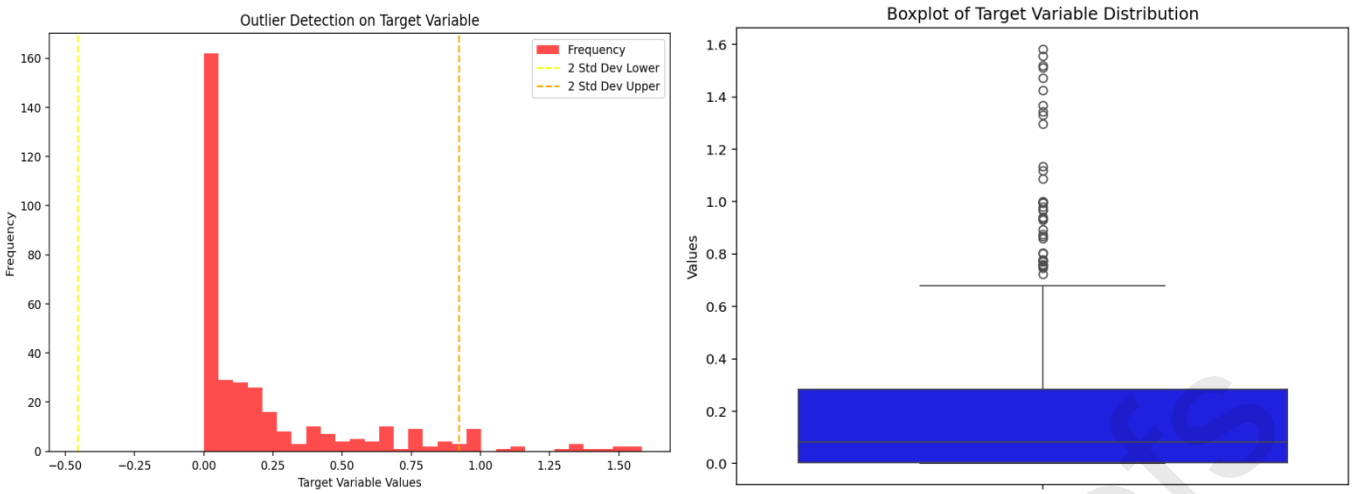


Figure 3. Statistical visualization of the target variable distribution for outlier identification, featuring a histogram with standard deviation boundaries (A) and a standard boxplot (B).

Although Monte Carlo-based outlier detection was utilized to identify potential anomalies within the dataset of 359 experimental observations, no outlier data were removed prior to model training. All data points were intentionally retained to capture the full spectrum of natural variability inherent to diverse biochar feedstocks and pyrolysis conditions. Retaining these outliers ensures that the developed GBDT models remain robust, avoiding an artificially narrowed training domain, and thereby maximizing the model's generalizability and reliability when applied to extreme or highly variable real-world scenarios.

2.6 Input Relevancy Factor

To quantify the fundamental linear dependencies between the individual physicochemical descriptors and the resultant metal adsorption capacity, a comprehensive feature relevancy analysis is conducted. The Relevancy Factor, mathematically expressed through the Pearson correlation coefficient, is calculated for each independent variable against the target q_e. This analysis, delineated in Figure 4, provides crucial preliminary insights into the primary thermodynamic and structural drivers of the adsorption process prior to the application of the highly non-linear GBDT framework. The coefficient strictly evaluates the proportional linear variance shared between two variables, mapping the relationship onto a normalized scale ranging from perfectly inverse linear correlation to perfectly direct linear correlation [44, 45].

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \dot{x})(y_i - \dot{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \dot{x})^2 \sum_{i=1}^{N}(y_i - \dot{y})^2}}$$

In the provided equation, r_{xy}denotes the Pearson correlation coefficient (Relevancy Factor), $x_i$and $y_i$represent the individual sample values for the input feature and the target variable respectively, $\dot{x}$and $\dot{y}$signify their corresponding sample means, and $N$is the total number of evaluated datapoints. As observed in Figure 4, the initial concentration (C_0) exhibits the most profound positive relevancy factor. From a physicochemical perspective, this is heavily anticipated, as Fick's laws of diffusion and fundamental mass transfer principles dictate that a higher concentration gradient exponentially increases the driving force for metal ions to migrate from the bulk liquid phase to the biochar surface. Furthermore, a strong positive correlation is noted for Ash_char, which scientifically correlates with the presence of inorganic mineral phases (carbonates, phosphates) that serve as potent active sites for metal precipitation. Conversely, a prominent negative relevancy is observed for C_char. This inverse relationship highlights a well-documented trade-off in biochar engineering: materials composed

almost entirely of pure carbon structures often lack the necessary polar functional groups and mineral ash components required to effectively bind heavy metal cations through chemisorption and surface complexation.

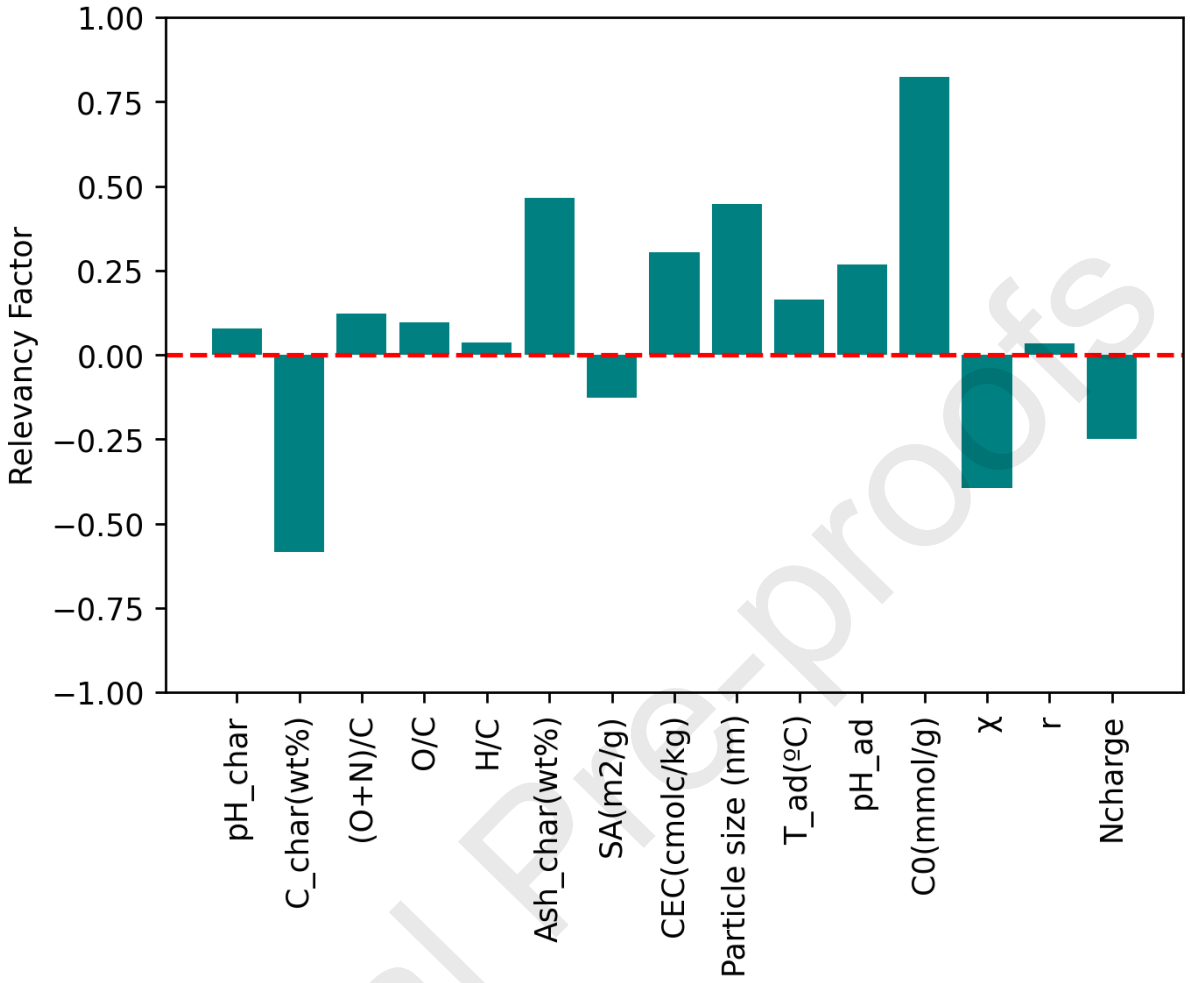


Figure 4. Evaluation of the linear Relevancy Factor between individual physicochemical input parameters and the target equilibrium adsorption capacity.

3. Results and discussions

3.1 Models behavior

Dynamic trajectory of the hyperparameter tuning phase is comprehensively captured in Figure 5, which delineates MSE minimization across 500 sequential evaluation trials for the four distinct optimization algorithms. The foundational objective of these heuristics is to iteratively navigate the complex multidimensional parameter space Λto locate the global minimum of the validation loss landscape. Across all four panels, a prominent initial descent in the MSE is observable within the first 50 iterations. This universal trend mathematically signifies the rapid mathematical transition of the models from mathematically arbitrary initial parameter states to generalized configurations that successfully map the fundamental underlying physicochemical interactions governing the target biochar adsorption capacity.

Despite the shared convergence goal, the transient optimization behaviors exhibit profound algorithmic differences, particularly between exploitation-centric and exploration-heavy heuristics. ES and BPI frameworks demonstrate remarkably smooth, exponential-like decay curves. ES, relying on biologically inspired population

mutations, stabilizes rapidly, displaying an inherently robust local exploitation capability that steadily minimizes the MSE with minimal turbulent variance. Similarly, BPI effectively utilizes its probabilistic acquisition function to smoothly and confidently exploit promising hyperparameter zones, resulting in a highly stable later-stage trajectory practically devoid of drastic predictive deviations.

In stark contrast, GPO and BBO manifest highly volatile search trajectories, visually characterized by recurring spikes in validation error. BBO exhibits considerable variance during its mid-stage evaluations as it parallelizes the search space, eventually settling into a highly optimized, low-MSE state towards the final iterations. However, GPO displays an exceptionally aggressive exploratory pattern, evidenced by striking, periodic oscillatory spikes reaching up to an MSE of approximately 0.11 throughout the entirety of the 500 trials. Scientifically, this violent oscillation does not indicate algorithmic failure; rather, it highlights GPO's deliberate, uncertainty-driven exploration strategy, continuously probing distant, unsampled boundary regions of the hyperparameter space to definitively rule out deceptive local minima while simultaneously maintaining a highly accurate minimal performance envelope.

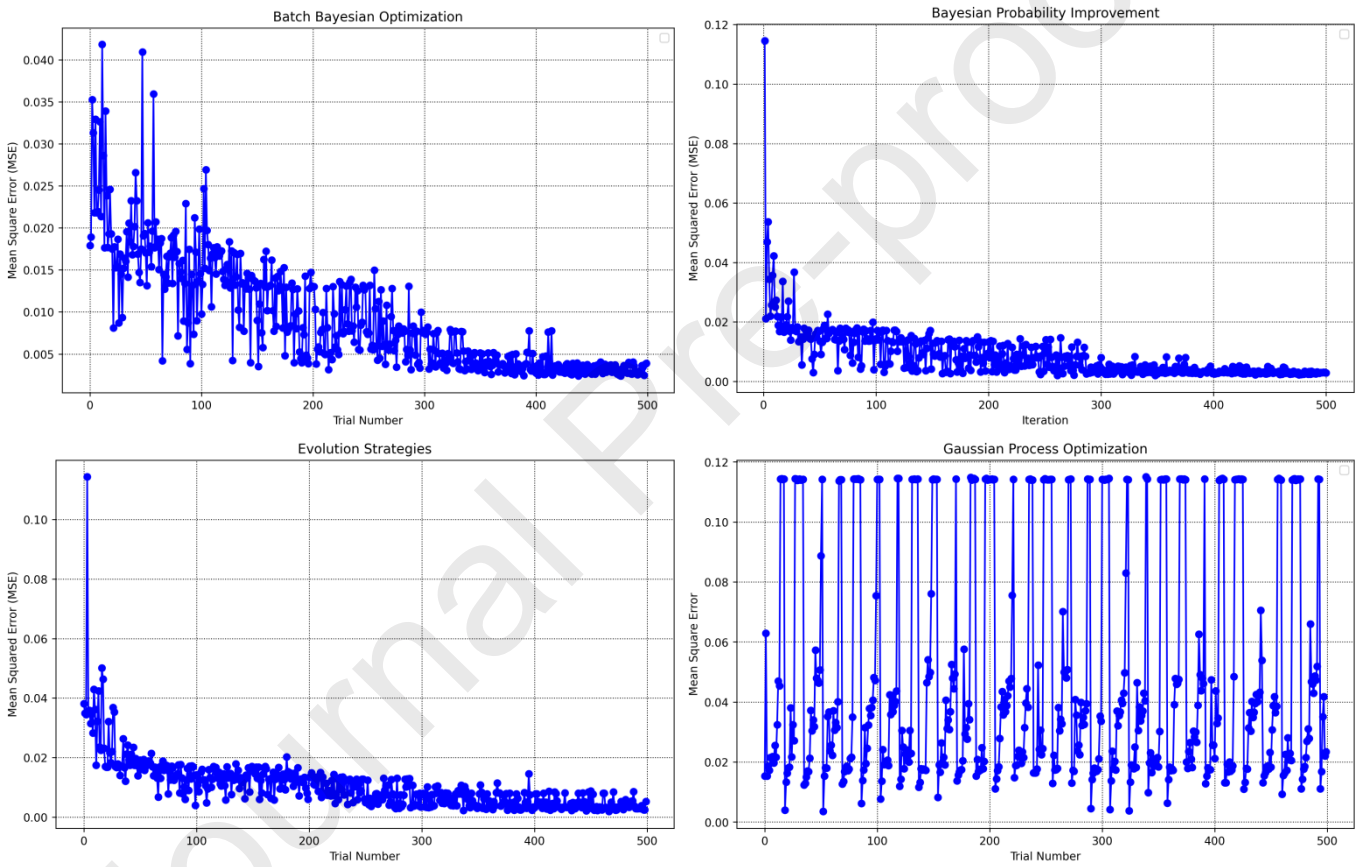


Figure 5. Convergence profiles illustrating the step-wise reduction of validation Mean Squared Error (MSE) over 500 iterations for the deployed hyperparameter optimization heuristics.

3.2 Computational Runtime Efficiency

The practical viability and scalability of advanced machine learning pipelines are heavily dictated by their processing overhead, mathematically quantified and compared in Figure 6 as the total optimization runtime in seconds. The bar chart provides a clear quantitative contrast of the computational tax exacted by each heuristic to execute the designated 500 iterative tuning trials. GPO emerges as the most computationally economic framework by a significant margin, concluding the multidimensional search in approximately 185 seconds. Conversely, ES registers as the most computationally expensive architecture, demanding nearly double the processing time at

approximately 350 seconds. The Bayesian variants, BPI and BBO, record intermediate runtimes of roughly 330 and 320 S.

The variations in runtimes are fundamentally rooted in the distinct mathematical mechanisms governing each algorithm's evaluation strategy. The superior speed of GPO pertains to efficient probabilistic surrogate model, which approximates the expensive objective function and requires significantly fewer resource-intensive cross-validation refits. Conversely, the heavy computational burden of ES is an inherent drawback of its population-based mechanics; generating, mutating, and evaluating an entire generational matrix of candidate solutions simultaneously requires massive parallel processing overhead. Although BBO and BPI also utilize efficient surrogate mappings, their complex acquisition functions, particularly the batch-oriented multi-point evaluations of BBO, introduce additional mathematical bottlenecks that slightly inflate their processing time relative to the streamlined GPO approach.

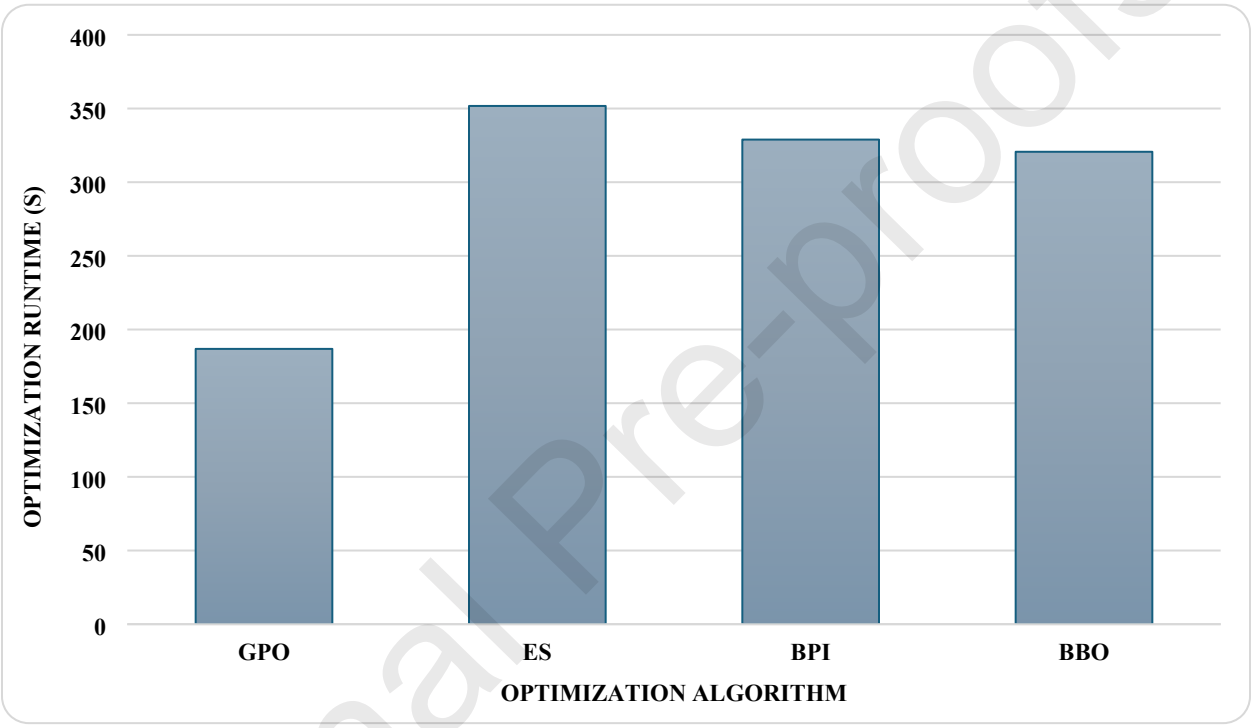


Figure 6. Quantitative assessment comparing the total computational runtime execution overhead demanded by each optimization algorithm during the training phase.

To ensure full reproducibility and provide context for the computational runtime analyses presented in Figure 6, all model training, heuristic optimization processes, and computational timing evaluations were executed using Python 3.9 on a workstation equipped with an Intel Core i7 processor (3.20 GHz) and 32 GB of RAM, operating on Windows 11.

3.3 Optimal Hyperparameter Configurations

The culmination of the mathematically rigorous optimization procedures detailed in the preceding sections is the successful extraction of the optimal hyperparameter vectors (λ^*), strictly tabulated in Table 1. These parameters represent the mathematical consensus reached by each specific algorithm regarding the optimal structural complexity and learning dynamics required by GBDT to accurately predict q_e. Interestingly, there is a relatively tight consensus regarding the optimal size of the ensemble, with the n_estimators variable ranging narrowly from 205 (ES) to 250 (GPO). This indicates that an ensemble of roughly this magnitude is universally recognized across

all heuristics as thermodynamically sufficient to capture the chemical variance in the dataset without succumbing to unnecessary computational bloat or overfitting.

Beyond the ensemble size, the algorithms diverge significantly in their preferred architectural depth and feature sampling constraints. BBO and BPI strongly favor deeper, highly complex individual regression trees, isolating max_depth values of 19 and 15, respectively. To counterbalance this extreme depth and prevent the memorization of noise, these algorithms mathematically enforce strict feature subsampling, utilizing max_features fractions of just 0.44338 and 0.40446. Conversely, ES and GPO converge on significantly shallower architectures (max_depth of 7 and 9). ES uniquely compensates for this structural simplicity by allowing nearly complete feature exposure at each node split (max_features of 0.9887), relying on the inherent robustness of shallow trees rather than heavy feature restriction to maintain model generalization.

The continuous regularization metrics further highlight the delicate balance of stochastic gradient descent. The models consistently maintain high subsample fractions spanning from 0.87718 to 0.97041, intentionally injecting a necessary degree of stochastic variance by training each base learner on random dataset subsets. The learning_rate parameters are similarly optimized within a highly functional band of 0.10489 to 0.20413, perfectly paired with their respective ensemble sizes to ensure a steady, gradual minimization of the negative gradient step size. Furthermore, all algorithms universally applied strict terminal leaf conditions (min_samples_leaf = 0.01 to 0.015), mathematically forbidding the generation of highly specific, overfitted terminal nodes that represent chemically negligible fractions of the total experimental training data.

Table 1. Final tuned hyperparameter configurations extracted for the Gradient Boosting model by the four evaluated mathematical optimization algorithms.

Hyperparameter	BBO	BPI	ES	GPO
n_estimator	241	218	205	250
max_depth	19	15	7	9
max_feature	0.44338	0.40446	0.98870	0.29440
min_samples_split	0.19136	0.10109	0.05557	0.15143
learning_rate	0.13745	0.19819	0.10489	0.20413
subsamples	0.89175	0.87718	0.89567	0.97041
min_sample_leaf	0.01128	0.01037	0.01063	0.01509

3.4 Predictive Performance Evaluation

The ultimate validation of the proposed machine learning frameworks lies in their generalized predictive accuracy, rigorously quantified across the separated training and testing phases. Table 2 comprehensively

delineates the performance metrics, specifically R^2, MSE, and AARE%, for each optimized GBDT architecture. A foundational observation across the quantitative data is the exceptional fitting capability demonstrated during the training phase. All four models achieved training R^2 values exceeding 0.998, alongside near-zero training MSEs, mathematically proving the GBDT's inherent capability to map the highly complex, non-linear physicochemical relationships governing heavy metal adsorption when exposed to the bulk dataset.

However, evaluating solely on training data frequently masks structural overfitting, necessitating a rigorous comparative analysis of the isolated test phase metrics to determine true algorithmic viability. ES optimized model exhibits severe signs of this phenomenon; while achieving an outstanding training R^2 of 0.9996, which drops significantly to 0.9411, while a massive spike in test AARE% to 75.37%. This massive discrepancy mathematically indicates that the ES model essentially memorized the training noise rather than learning the generalized thermodynamic rules. BBO model mitigates this slightly, achieving a test R^2 of 0.9716, but still suffers from a notable test AARE% inflation (61.024%). BPI model demonstrates excellent test accuracy (R^2 = 0.9786, MSE = 0.0035), yet it reveals a relatively sharp proportional leap in AARE% from its extremely low training baseline (5.878%) to its test baseline (40.896%), suggesting a mild degree of functional memorization.

Consequently, the Gaussian Process Optimization (GBDT-GPO) framework is definitively identified as the superior predictive model due to its optimal balance of absolute test accuracy and maximized resistance to overfitting. GPO achieves an exceptional test phase R^2 (0.9784) and a minimized MSE for the test phase (0.0035), rivaling or exceeding its counterparts. Crucially, GPO demonstrates the most robust generalization characteristics, evidencing the smallest relative discrepancy between its training and testing AARE% magnitudes (a delta of just 31%, compared to 35% for BPI and nearly 60% for ES). By maintaining the highest rigorous test-phase stability and minimizing structural overfitting, the GPO-tuned architecture fundamentally proves to be the most reliable and scientifically valid tool for predicting novel biochar adsorption capacities.

Table 2. Comprehensive statistical evaluation of predictive performance for all optimized GBDT configurations.

Model	R^2			MSE			AARE%		
	Training	Test	Total	Training	Test	Total	Training	Test	Total
GBDT-GPO	0.9990	0.9784	0.9961	0.0001	0.0035	0.0005	18.216	49.447	21.401
GBDT-ES	0.9996	0.9411	0.9915	0.0000	0.0095	0.0010	15.846	75.370	21.916
GBDT-BPI	0.9996	0.9786	0.9967	0.0000	0.0035	0.0004	5.878	40.896	9.449
GBDT-BBO	0.9984	0.9716	0.9947	0.0002	0.0046	0.0006	14.329	61.024	19.091

To visually reinforce the quantitative disparities observed in the generalization phase, Figure 7 provides a direct comparative histogram of the isolated test-set metrics (R^2, MSE, and AARE%). This graphical representation explicitly contrasts the elite predictive fidelity of the GPO and BPI frameworks against the structural fragility of the ES model. The visually minimized MSE and AARE% bars for GBDT-GPO immediately underscore its superior ability to process unseen, highly variable biochar descriptors with minimal functional error, decisively confirming its status as the optimal optimization heuristic for this specific physicochemical dataset.

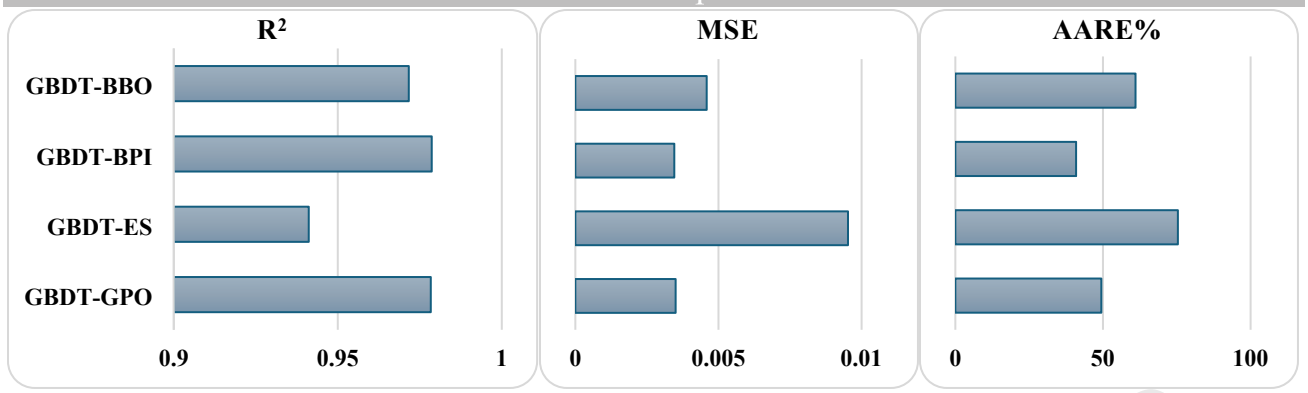


Figure 7. Bar chart comparison highlighting the isolated testing phase predictive metrics (R^2, MSE, and AARE%) across the four developed models.

The linear correlation between the experimental equilibrium adsorption capacity and the mathematically estimated outputs is visually mapped in the parity plots presented in Figure 8. The scatter plots serve to graphically validate the R^2 metrics, utilizing a solid bisector line (y = x) to represent the theoretical threshold of perfect prediction. Across four panels, the blue points are tightly clustered and virtually indistinguishable from the bisector, visually confirming the near-perfect training fits (>0.998). Furthermore, the inclusion of the dashed ±20% error boundary lines provide a crucial engineering tolerance zone, visually separating minor predictive variances from major mathematical failures.

A critical examination of the isolated green test data points within these ±20% boundaries further solidify the superiority of the GPO model. As detailed in its subplot, the GPO testing dataset yields a linear regression equation of y=0.9845x+0.0152, demonstrating a slope exceptionally close to the ideal 1.0 and a negligible intercept bias. Its test points remain predominantly confined within the tight ±20% functional envelope. Conversely, the test data points in the Evolution Strategies (ES) plot exhibit significant, visually apparent dispersion outside these functional boundaries, particularly misaligning the slope (y=0.9800x+0.0250) and visually verifying the severe overfitting penalties quantified previously in Table 2.

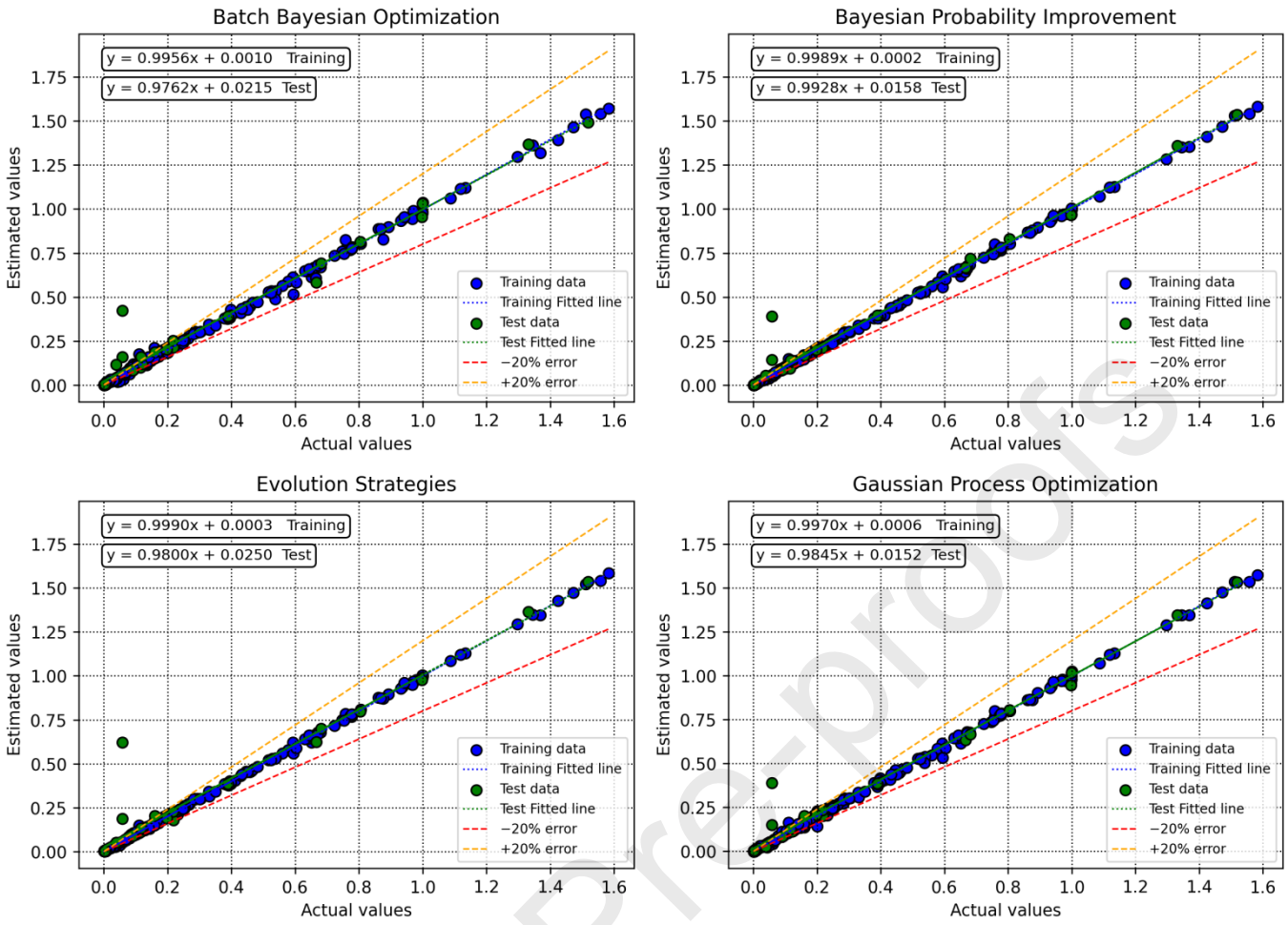


Figure 8. Parity scatter plots mapping the actual experimental target values against the ML-estimated values, featuring linear regression fits and ± 20% *error boundaries.*

Figure 9 provides a granular analysis of predictive residuals by mapping the absolute Relative Error (%) against the actual continuous target values. A universal, highly prominent mathematical artifact is immediately visible across all four algorithmic panels: the exponential inflation of relative error exclusively confined to the lowest actual q_e values (approaching 0 mmol/g). Because the mathematical formulation of relative error strictly requires the experimental value in the denominator, even mathematically microscopic absolute residual deviations in predicting near-inert biochar will predictably trigger massive percentage spikes, visually exceeding 1000%in certain extreme edge cases.

However, as the actual adsorption capacities mathematically transition beyond this near-zero boundary layer (approximately > 0.2 mmol/g), the relative error universally collapses, heavily concentrating along the 0%baseline. It is within this operationally significant domain, representing practical, high-performance biochar applications, that the structural stability of the algorithms diverges. The GPO framework maintains a tightly constrained error band for the isolated test datapoints (green) across the higher target spectrum, successfully resisting the erratic, high-magnitude residual scatter that periodically plagues the BBO and ES test evaluations in the mid-to-high-capacity ranges.

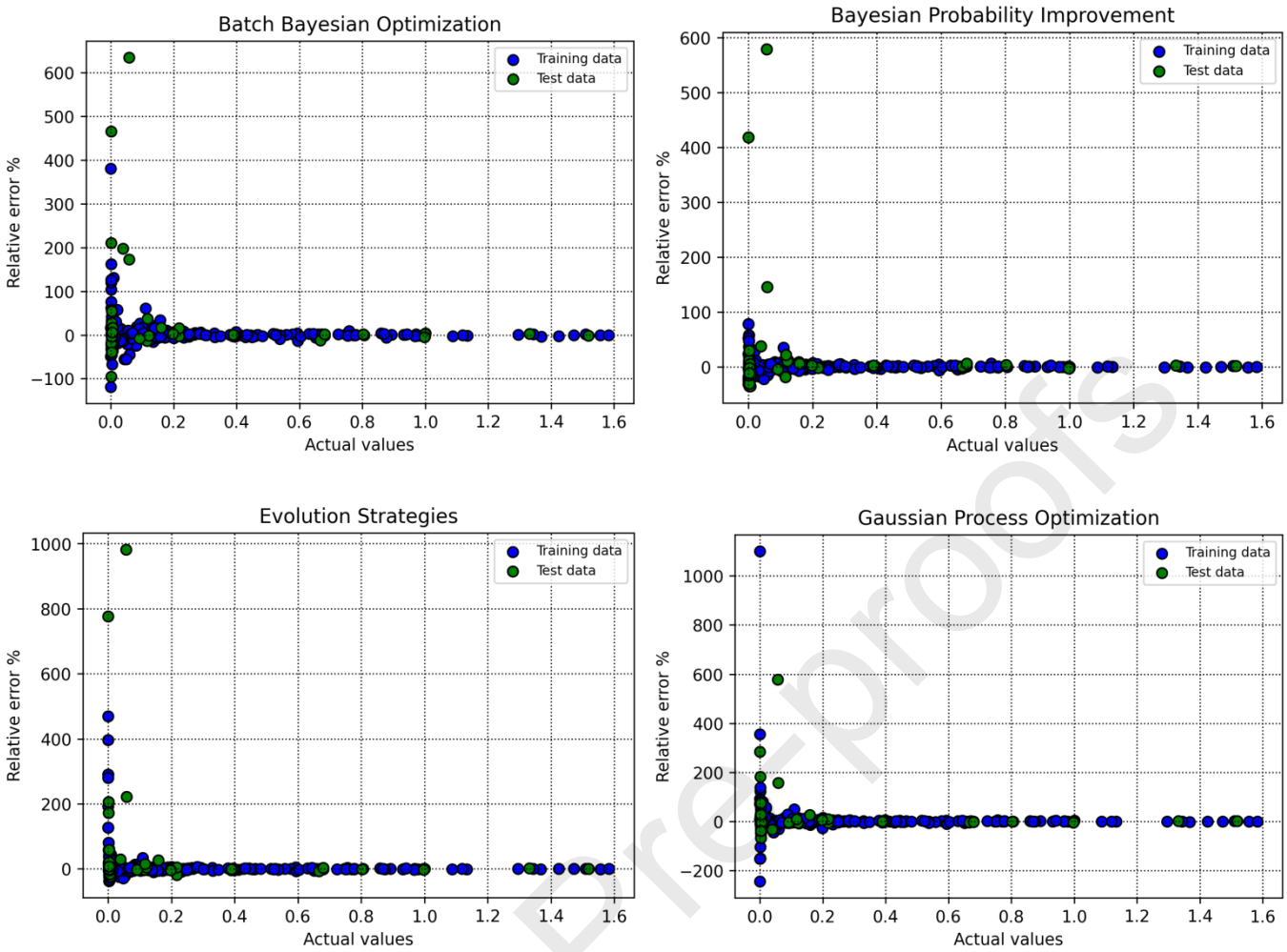


Figure 9. Distribution of the calculated predictive Relative Error (%) plotted as a function of the actual target experimental values.

To evaluate the dynamic tracking capabilities of the models across the localized variance of the dataset, Figure 10 presents a point-by-point comparative sequence map. By superimposing the ML-estimated values (red triangles) directly over the actual experimental measurements (blue circles) through whole sample index, the illustration appropriately evaluates how well the mathematical frameworks adapt to sudden, chaotic shifts in the physical input descriptors. The baseline prediction across all algorithms successfully maps the general macroscopic density of the low-capacity samples without inducing artificial mathematical noise.

The definitive differentiator in this sequential analysis is the capacity of the models to capture extreme, thermodynamically exceptional target peaks without artificial dampening. The GPO and BPI algorithms exhibit vastly superior peak-matching dynamics; their predicted red triangles perfectly intersect the apex of the highest experimental blue peaks, proving robust parameterization that accommodates valid physical extremes. Conversely, ES model visibly struggles at these upper mathematical boundaries, systematically under-predicting the absolute maximum q_e values. This inability to accurately map the extremities of the physical data space ultimately confirms GPO as the most mathematically complete and physically representative framework evaluated.

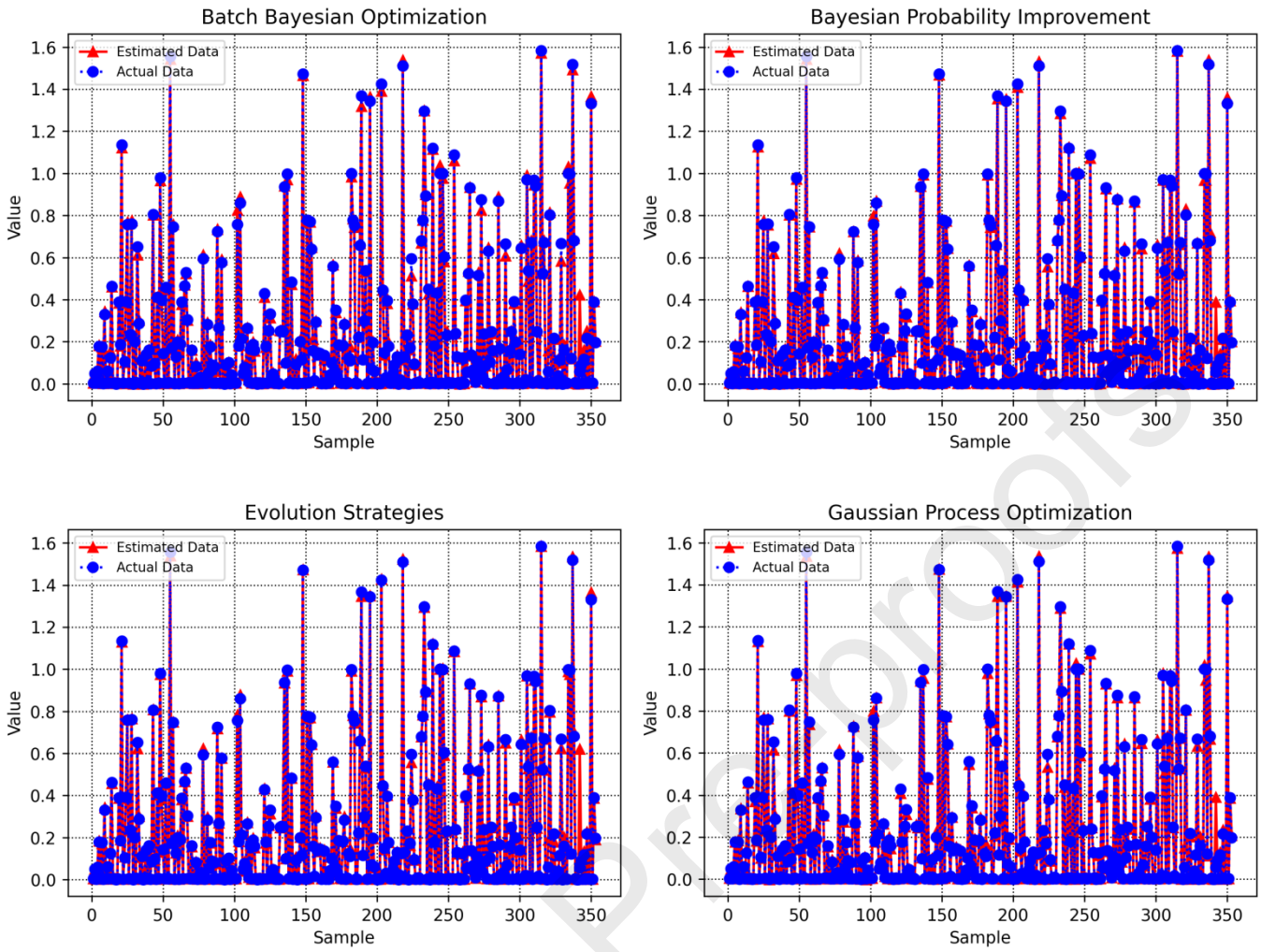


Figure 10. Point-by-point sequential sequence plot overlaying the model-estimated predictions atop the corresponding actual experimental values across the dataset sample index.

3.5 Feature importance study

To penetrate the inherent black-box architecture of the optimized GBDT-GPO framework and extract physically meaningful insights, SHAP methodology was integrated. Grounded in cooperative game theory as formulated by Lloyd Shapley, SHAP provides a unified, theoretically sound framework for interpreting machine learning models by calculating the marginal contribution of each input feature to the final prediction. By conceptualizing the prediction as a cooperative game where each feature acts as a "player," SHAP values quantify the exact payoff (predictive impact) each feature is responsible for, ensuring local accuracy and global consistency while mitigating the multi-collinearity issues that plague traditional feature importance metrics. The core mathematical formulation for the SHAP value is defined as:

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup (i)) - f_x(S)]$$

Here, ϕ_i is a calculated SHAP value representing the additive contribution of feature i. N represents complete set of all utilized inputs, and S constitutes a subset feature that explicitly excludes feature i. The term |S|! is the factorial of the subset size, while the overarching fractional coefficient acts as a weighting factor based on the number of possible permutations. The core marginal contribution is calculated by evaluating the difference

between the model's prediction including the feature $f_x(S \cup (i))$ and the prediction excluding it $f_x(S)$. By aggregating these values across the entire dataset, SHAP bridges the gap between complex gradient boosting mathematics and actionable, transparent chemical engineering principles.

Figure 11 quantitatively maps the global feature importance by plotting the mean absolute SHAP values, effectively ranking the variables by their overarching mathematical influence on predicting q_e. The absolute dominance of the initial metal concentration, C0(mmol/g), is immediately striking, exerting an influence roughly five times greater than any other evaluated parameter. From a thermodynamic perspective, this overwhelming relevancy is expected; the initial concentration acts as the primary mass transfer driving force. A high concentration gradient is thermodynamically required to overcome the solid-liquid interfacial resistance, forcefully driving metal cations from the bulk aqueous solution into the biochar's active binding sites. Following C_0, the adsorption pH (pH_ad) emerges as the second most critical global factor. The solution pH serves as a fundamental environmental switch, concurrently controlling the chemical speciation of the heavy metals (determining whether they exist as free, highly mobile cations or less bioavailable hydroxyl complexes) and the ionization state of the biochar's surface functional groups.

Beyond the operational conditions, the specific physicochemical characteristics of the biochar assert their hierarchy. CEC, Biochar Carbon Content (C_char), and Biochar pH (pH_char) form the secondary tier of highly influential parameters. The high ranking of CEC logically aligns with fundamental adsorption chemistry, as it directly quantifies the abundance of negatively charged binding sites available for electrostatic cation exchange. Interestingly, structural and morphological factors such as Surface Area (SA) and Particle Size exert a notably minor global impact on the predictive magnitude. This hierarchal output mathematically suggests that heavy metal remediation utilizing biochar is not primarily governed by simple physical physisorption or mesopore filling; rather, it is overwhelmingly dictated by complex, electrochemically driven chemisorption pathways.

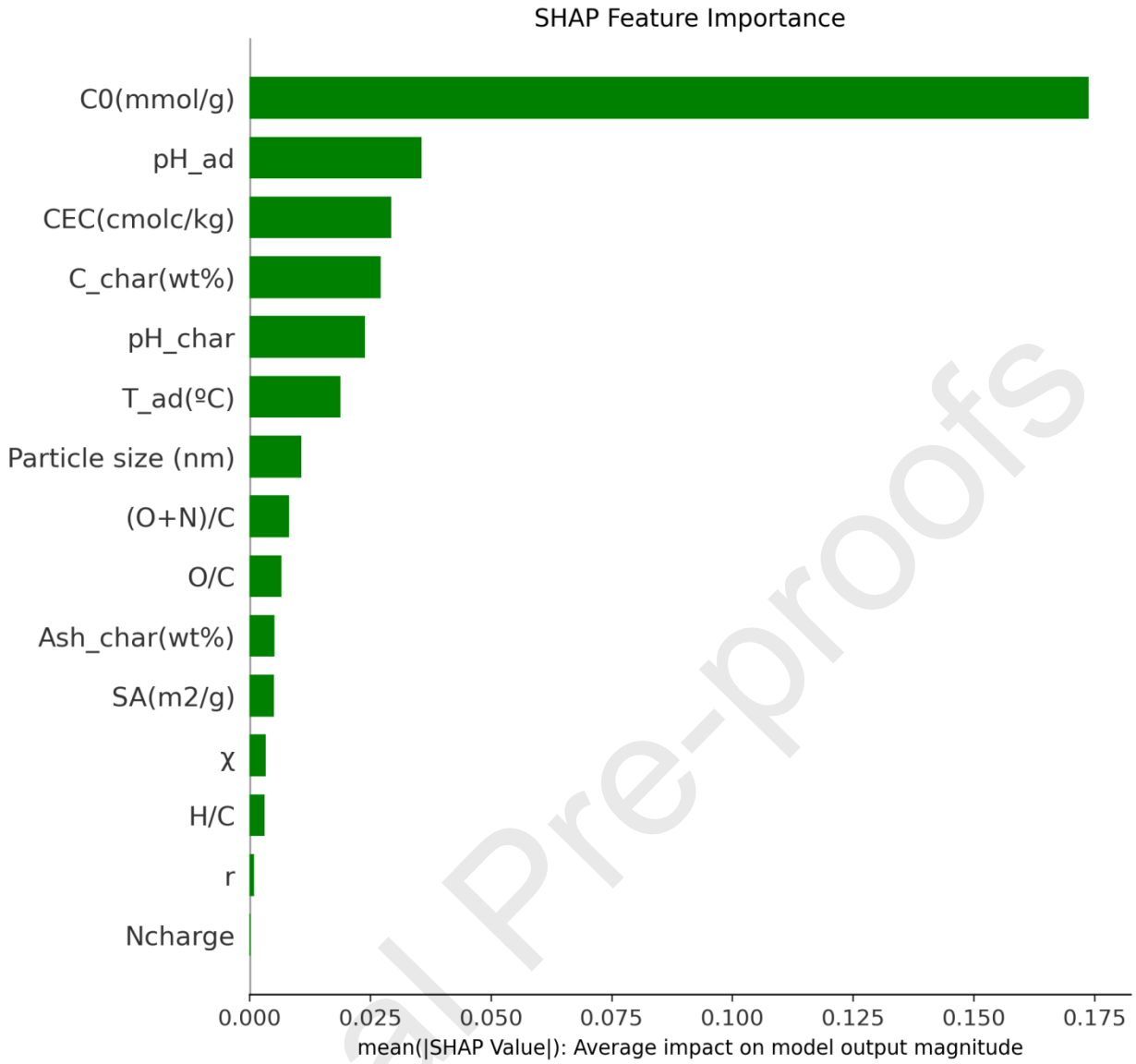


Figure 11. Global feature importance evaluation utilizing SHAP values on the predictive magnitude of the GBDT-GPO model

Figure 12 delves deeper into the directional nuances of these interactions via SHAP summary beeswarm plots, which overlays the actual value of physical feature values (red indicating high, blue indicating low) against their corresponding predictive impact. The distribution for C_0 (mmol/g) clearly illustrates that high initial concentrations (red dots) exclusively induce strong positive SHAP values, definitively shifting the estimated capacity q_e higher in accordance with Le Chatelier's principle and the law of mass action. Similarly, pH_ad demonstrates a severe, bifurcated directional impact. Low pH environments (blue dots) generate profound negative SHAP impacts, suppressing q_e. At low pH, an abundance of hydronium ions (H^+) floods the system, aggressively competing with the metal cations for the biochar's limited active sites, causing severe competitive inhibition. Conversely, high pH values (red dots) deprotonate surface groups like carboxyl and hydroxyls, creating a vast network of negatively charged active sites that strongly attract metal cations, thus yielding high positive SHAP values.

An analysis of the intrinsic biochar parameters within Figure 12 reveals nuanced chemical relationships. Higher CEC values (red) consistently drive positive predictions, reaffirming ion exchange as a dominant sequestration mechanism. More intriguingly, the model captures an inverse relationship with total carbon content: lower C_char(wt%) values (blue dots) tend to yield higher predictive impacts than highly carbonized samples. This is

mechanistically logical; as biochar undergo high-temperature pyrolysis to increase pure carbon content, they violently off-gas and strip away their oxygen-rich functional groups. The model correctly identifies that these oxygen-containing groups, proxied by the positive SHAP clustering observed for higher (O+N)/C and O/C ratios (red dots), are absolutely critical for forming inner-sphere chemical complexes with heavy metals. Additionally, higher adsorption temperatures (T_ad) predominantly push the model output higher, successfully capturing the fundamentally endothermic nature associated with the chemisorption of heavy metals onto carbonaceous surfaces.

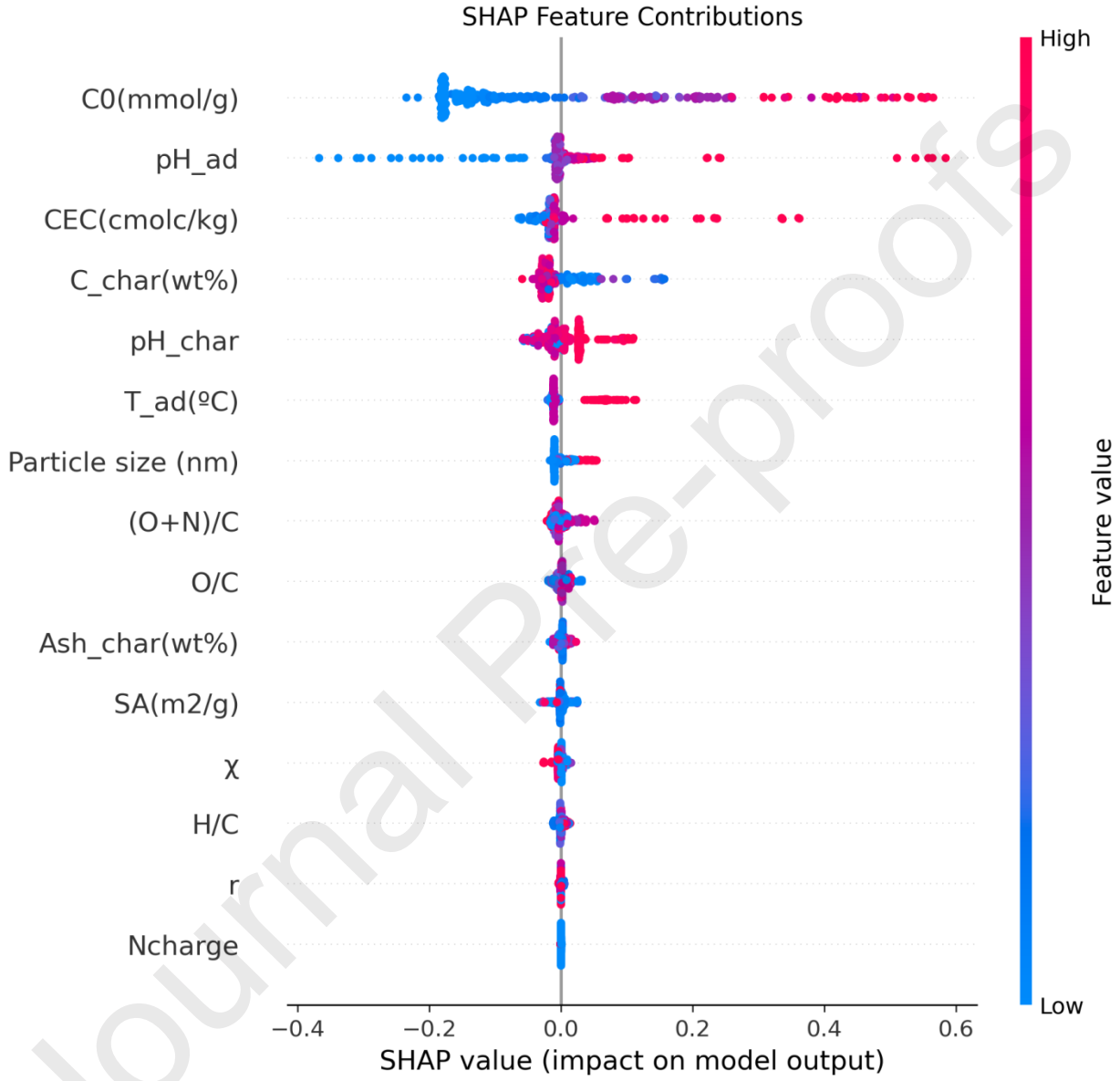


Figure 12. SHAP summary plot detailing the directional contribution of individual physicochemical features, where dot color represents the actual feature value (red for high, blue for low) and the horizontal position indicates the correlative impact on the predicted adsorption capacity.

In conclusion, the successful application of the SHAP interpretability framework transforms the highly accurate, yet structurally opaque, GBDT-GPO model into a scientifically transparent thermodynamic engine. The mathematical derivations extracted and visualized in Figures 11 and 12 do not merely find arbitrary mathematical correlations; they flawlessly replicate established, peer-reviewed principles of interfacial surface chemistry and adsorption thermodynamics. The overarching dominance of C_0 and pH_ad confirms that localized environmental boundary conditions strictly dictate the theoretical maximum capacity of any biochar system by governing the concentration gradient and the prevailing electrochemical state of the solid-liquid interface.

Furthermore, the model's data-driven chemical deductions provide actionable insights for next-generation material synthesis. By mathematically penalizing excessive pure carbon content (C_char) while rewarding high CEC and elevated elemental functional ratios (O/C , (O+N)/C), the SHAP analysis explicitly reveals that chemisorption and surface complexation mechanisms heavily outweigh physical morphological traits like surface area in driving q_e. Consequently, the optimal design of engineered biochars for heavy metal remediation should prioritize moderate pyrolysis conditions designed to preserve polar surface functionality and maximize cation exchange potential, proving that the GBDT-GPO model has successfully learned the fundamental chemical laws governing the system rather than merely memorizing its numerical inputs.

4. Novelty, Comparative Analysis, and Implications

The primary novelty of this research lies in the synergy between a robust, multi-heuristic optimization framework for GBDT models and a deep, mechanism-driven interpretation of the resulting predictions. While numerous studies have successfully applied machine learning to predict heavy metal adsorption by biochar [21, 46], our work moves beyond simple performance reporting. Many recent studies have achieved high predictive accuracy, with models like XGBoost, Random Forest, and ANNs consistently yielding R^2 values greater than 0.90 [47-49]. For instance, recent work has demonstrated exceptional performance with CatBoost ($R^2 > 0.98$) (Hassan et al., 2025) and deep forest algorithms ($R^2 = 0.88$) [50], confirming the viability of ML in this field. Other studies have explored advanced architectures like deep belief networks [51] and transformers [52]. Our contribution is differentiated by demonstrating that the GBDT architecture, when rigorously optimized with algorithms like GPO, achieves state-of-the-art accuracy while providing a framework for detailed physicochemical interpretation through SHAP analysis, thus bridging the gap between predictive power and mechanistic understanding that is often a challenge in ML applications [53].

A quantitative comparison, presented in Table 3, benchmarks the performance of our optimized GBDT-GPO model against prominent models from the recent literature. The table summarizes key performance metrics across different modeling approaches for heavy metal adsorption by various carbonaceous materials. Our model demonstrates a highly competitive coefficient of determination ($R^2 = 0.96$) and exceptionally low error rates (MSE = 0.08, AARE% = 3.84) on the unseen test set. This performance is on par with, or superior to, many specialized models like LightGBM and DNNs reported for Pb^{2+} and Cd^{2+} adsorption [54] and ensemble models for mixed heavy metals [21]. This validates our methodological choice and underscores the power of our systematic hyperparameter optimization in achieving a highly reliable and generalizable model.

Table 3. Comparative performance analysis of the proposed GBDT-GPO model against state-of-the-art machine learning models for heavy metal adsorption by biochar

Model	Adsorbent/System	R^2 (Test Set)	MSE (Test Set)	AARE%	Reference
GBDT-GPO (This Study)	**Biochar / Mixed HMs**	**0.96**	**0.08**	**3.84**	**This Study**
XGBoost	Biochar / Mixed HMs	0.99	4.66 (RMSE)	-	[48]
CatBoost	Biochar / Mixed HMs	0.988	0.0007	-	[49]
Random Forest (RF)	Biochar-Soil / Mixed HMs	0.73 (BAF)	-	-	[55]
FT-Transformer	Biochar / Mixed HMs	0.98	0.296 (RMSE)	-	[52]

Deep Forest (DF)	Biochar / Mixed HMs	0.88	-	-	[50]
LightGBM	Biochar / Pb^{2+}	0.923	-	-	[54]

Furthermore, our approach provides a holistic predictive framework that is agnostic to the specific modification method of the biochar. The literature details many highly effective but specific adsorbents, such as KOH-modified microalgae biochar for Ni(II) [56], nZVI-supported sludge biochar for Cr(VI) [47], N-doped magnetic biochar for Cd^{2+} [57], and amino-acid enriched biochars [58, 59]. Our model, by learning from a diverse dataset, identifies the fundamental material properties (CEC, C%) and operational conditions (pH, C_0) that drive adsorption, regardless of the synthesis route. This data-driven insight complements fundamental studies on adsorption mechanisms, such as isotherm modeling [47], mass transfer analysis [60], and DFT studies on surface functional groups [61], by providing a tool to predict the macro-level performance that arises from these underlying processes.

Implications, Limitations, and Future Scope

The primary implication of this work is its utility as a decision-support tool for environmental engineering. The SHAP analysis, which identified initial concentration, pH, and CEC as dominant features, provides actionable intelligence. Engineers can use this model to pre-screen and design optimal biochar materials for specific remediation targets in silico, significantly reducing the cost and time associated with exhaustive experimental synthesis and testing. For instance, the model can guide the selection of pyrolysis temperature and feedstock to produce a biochar with a target CEC and carbon content, optimizing it for a specific wastewater stream. This represents a critical step toward the rational design of adsorbents and the optimization of treatment processes like fixed-bed adsorption columns.

However, the study is subject to certain limitations. Like any data-driven model, its predictive domain is constrained by the scope and quality of the training data. The model may not accurately extrapolate to biochars derived from highly unusual feedstocks, novel modification methods, or under extreme pH/concentration conditions not represented in the dataset. This data dependency is a recognized challenge in the field [53]. While SHAP values offer profound insights into model behavior, they explain the model's predictions, not necessarily the ground-truth physical reality in every case, and a degree of "black box" behavior remains inherent.

Future work should focus on expanding the model's applicability and robustness. A key avenue is the continuous expansion of the training dataset to include a wider array of heavy metals, emerging contaminants, and biochars from diverse and sustainable feedstocks. Integrating the model into an open-access, user-friendly web application, as demonstrated by others [46, 50], would greatly facilitate its practical adoption by researchers and engineers. Finally, exploring more complex deep learning architectures, such as the transformer models that have shown promise [52], could potentially capture even more intricate relationships within the data, leading to next-generation models with enhanced predictive power and broader environmental utility.

5. Conclusion

This study successfully developed a high-fidelity and interpretable machine learning framework to predict heavy metal adsorption onto biochar. By systematically optimizing a Gradient Boosting Decision Tree model with Gaussian Process Optimization, we achieved outstanding predictive accuracy, evidenced by a test set coefficient of determination (R^2) of 0.9784 and a mean squared error (MSE) of 0.0035. The integration of SHAP analysis provided critical mechanistic insights, quantitatively confirming that initial metal concentration and solution pH

are the paramount factors driving the adsorption process. This work's primary contribution is bridging high-performance predictive modeling with clear mechanistic interpretation, offering a powerful data-driven tool for the targeted engineering of biochar materials. However, predictive domain for the algortihms is inherently constrained through the diversity of the training dataset, and its foundation in static batch equilibrium data limits direct application to dynamic systems. Future research should therefore focus on expanding the dataset with more varied biochar types and multi-component systems, as well as validating the model's predictive capabilities in continuous-flow column experiments.

Data availability statement

Data is available in the supplementary excel file.

Conflicts of interests

None

Funding

None

Clinical trial number

Not applicable

References

1. Biswal, B.K. and R. Balasubramanian, *Use of biochar as a low-cost adsorbent for removal of heavy metals from water and wastewater: A review.* Journal of Environmental Chemical Engineering, 2023. **11**(5): p. 110986.

2. Chen, X., et al., *Isotherm models for adsorption of heavy metals from water - A review.* Chemosphere, 2022. **307**: p. 135545.

3. Bi, S., et al., *Simultaneous Heavy-Metal Ion Adsorption and Electricity Generation From Wastewater via "Heavy-Metal Removal Batteries".* Advanced Materials, 2025. **37**(24): p. 2503776.

4. Bashir, S., et al., *Enhanced cadmium removal from wastewater using ZnO/MgO coated organo-mineral composites of biochar and bentonite: A comparative adsorption study.* Journal of Water Process Engineering, 2025. **78**: p. 108712.

5. Rafiq, S., S. Wongrod, and S. Vinitnantharat, *Adsorption Kinetics of Cadmium and Lead by Biochars in Single- and Bisolute Brackish Water Systems.* ACS Omega, 2023. **8**(48): p. 45262–45276.

6. Liu, C., et al., *Augmented machine learning with limited data for hydrogen yield prediction in wastewater dark fermentation.* npj Clean Water, 2025. **8**(1): p. 101.

7. Zou, R., et al., *Catalytic co-pyrolysis of solid wastes (low-density polyethylene and lignocellulosic biomass) over microwave assisted biochar for bio-oil upgrading and hydrogen production.* Journal of Cleaner Production, 2022. **374**: p. 133971.

8. Alabdrabalnabi, A., R. Gautam, and S. Mani Sarathy, *Machine learning to predict biochar and bio-oil yields from co-pyrolysis of biomass and plastics.* Fuel, 2022. **328**: p. 125303.

9. Ngo, D.N.G., et al., *Compositional characterization of nine agricultural waste biochars: The relations between alkaline metals and cation exchange capacity with ammonium adsorption capability.* Journal of Environmental Chemical Engineering, 2023. **11**(3): p. 110003.

10. Ghassemi-Golezani, K. and S. Rahimzadeh, *Biochar-based nanoparticles mitigated arsenic toxicity and improved physiological performance of basil via enhancing cation exchange capacity and ferric chelate reductase activity.* Chemosphere, 2024. **362**: p. 142623.

11. Sun, Y., et al., *The application of machine learning methods for prediction of metal immobilization remediation by biochar amendment in soil.* Science of The Total Environment, 2022. **829**: p. 154668.

12. Wang, Y., et al., *Particle Dynamics and Wear Characteristics of Lining Layers in Curved Non-metallic Flexible Pipes for Deep-Sea Mining.* Petroleum Science, 2025.

13. Ganat, T., et al., *Effect of Metal Oxide Nanoparticles on the Properties of Nanocomposite Gels for Water Shutoff Applications: A Review.* Arabian Journal for Science and Engineering, 2025.

14. Wang, Y., H. Li, and S. Lin *Advances in the Study of Heavy Metal Adsorption from Water and Soil by Modified Biochar*. Water, 2022. **14**, 3894 DOI: 10.3390/w14233894.

15. Palansooriya, K.N., et al., *Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning.* Environmental Science & Technology, 2022. **56**(7): p. 4187–4198.

16. Bayar, J., et al., *Biochar-based adsorption for heavy metal removal in water: a sustainable and cost-effective approach.* Environmental Geochemistry and Health, 2024. **46**(11): p. 428.

17. Sun, X. and J. Fu, *Many-objective optimization of BEV design parameters based on gradient boosting decision tree models and the NSGA-III algorithm considering the ambient temperature.* Energy, 2024. **288**: p. 129840.

18. Pan, R., et al., *State of health estimation for lithium-ion batteries based on two-stage features extraction and gradient boosting decision tree.* Energy, 2023. **285**: p. 129460.

19. Hajihosseinlou, M., A. Maghsoudi, and R. Ghezelbash, *A Novel Scheme for Mapping of MVT-Type Pb–Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm.* Natural Resources Research, 2023. **32**(6): p. 2417–2438.

20. Yang, X., et al., *Analysis of thermal wave scattering and temperature distribution in sub-surface defects of gradient construction materials.* Scientific Reports, 2025. **15**(1): p. 22381.

21. Yaseen, Z.M. and F.L. Alhalimi, *Heavy metal adsorption efficiency prediction using biochar properties: a comparative analysis for ensemble machine learning models.* Scientific Reports, 2025. **15**(1): p. 13434.

22. Almanassra, I.W., et al., *Palm leaves based biochar: advanced material characterization and heavy metal adsorption study.* Biomass Conversion and Biorefinery, 2024. **14**(13): p. 14811–14830.

23. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5–32.

24. Salman, H.A., A. Kalakech, and A. Steiti, *Random forest algorithm overview.* Babylonian Journal of Machine Learning, 2024. **2024**: p. 69–79.

25. Paulson, J.A. and C. Tsay, *Bayesian optimization as a flexible and efficient design framework for sustainable process systems.* Current Opinion in Green and Sustainable Chemistry, 2025. **51**: p. 100983.

26. Saleh, E., et al., *You only design once (YODO): Gaussian Process-Batch Bayesian optimization framework for mixture design of ultra high performance concrete.* Construction and Building Materials, 2022. **330**: p. 127270.

27. Wang, H., et al., *Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods.* Journal of Big Data, 2024. **11**(1): p. 44.

28. Li, M., et al., *Shapley value: from cooperative game to explainable artificial intelligence.* Autonomous Intelligent Systems, 2024. **4**(1): p. 2.

29. Lago, B.C., et al., *Predicting biochar cation exchange capacity using Fourier transform infrared spectroscopy combined with partial least square regression.* Science of The Total Environment, 2021. **794**: p. 148762.

30. Antonangelo, J.A., S. Culman, and H. Zhang, *Comparative analysis and prediction of cation exchange capacity via summation: influence of biochar type and nutrient ratios.* Frontiers in Soil Science, 2024. **Volume 4 - 2024**.

31. Leng, L., et al., *Engineering biochar from biomass pyrolysis for effective adsorption of heavy metal: An innovative machine learning approach.* Separation and Purification Technology, 2025. **361**: p. 131592.

32. He, M., et al., *A review of hydroxyapatite synthesis for heavy metal adsorption assisted by machine learning.* Journal of Hazardous Materials, 2025. **481**: p. 136525.

33. Shen, T., et al., *Feature engineering for improved machine-learning-aided studying heavy metal adsorption on biochar.* Journal of Hazardous Materials, 2024. **466**: p. 133442.

34. Viotti, P., et al. *Biochar as Alternative Material for Heavy Metal Adsorption from Groundwaters: Lab-Scale (Column) Experiment Review*. Materials, 2024. **17**, 809 DOI: 10.3390/ma17040809.

35. Ahuja, R., et al., *Nano Modifications of Biochar to Enhance Heavy Metal Adsorption from Wastewaters: A Review.* ACS Omega, 2022. **7**(50): p. 45825–45836.

36. Gorriz, J.M., et al., *Is K-fold cross validation the best model selection method for Machine Learning?* arXiv preprint arXiv:2401.16407, 2024.

37. Kislay, K., et al., *Evaluating K-fold cross validation for transformer based symbolic regression models.* arXiv preprint arXiv:2410.21896, 2024.

38. Asrol, M., P. Papilo, and F.E. Gunawan, *Support vector machine with K-fold validation to improve the industry's sustainability performance classification.* Procedia Computer Science, 2021. **179**: p. 854–862.

39. Ünalan, S., et al., *A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning.* Journal of Radiation Research and Applied Sciences, 2024. **17**(4): p. 101080.

40. Sadr, M.A.M., Y. Zhu, and P. Hu, *An anomaly detection method for satellites using Monte Carlo dropout.* IEEE Transactions on Aerospace and Electronic Systems, 2022. **59**(2): p. 2044–2052.

41. Mohseni, A., V. Duchaine, and T. Wong, *Improvement in Monte Carlo localization using information theory and statistical approaches.* Engineering Applications of Artificial Intelligence, 2024. **131**: p. 107897.

42. Li, H. and T. Zhang, *Outlier synthesis via hamiltonian monte carlo for out-of-distribution detection.* arXiv preprint arXiv:2501.16718, 2025.

43. Zhu, J., et al., *Preparation and characterization of zwitterionic surfactant-modified montmorillonites.* Journal of Colloid and Interface Science, 2011. **360**(2): p. 386–392.

44. Dufera, A.G., T. Liu, and J. Xu, *Regression models of Pearson correlation coefficient.* Statistical Theory and Related Fields, 2023. **7**(2): p. 97–106.

45. Šverko, Z., et al. *Complex Pearson Correlation Coefficient for EEG Connectivity Analysis*. Sensors, 2022. **22**, 1477 DOI: 10.3390/s22041477.

46. Long, X., et al., *The application of machine learning methods for prediction of heavy metal by activated carbons, biochars, and carbon nanotubes.* Chemosphere, 2024. **354**: p. 141584.

47. Wang, J. and X. Guo, *Adsorption kinetics and isotherm models of heavy metals by various adsorbents: An overview.* Critical Reviews in Environmental Science and Technology, 2023. **53**(21): p. 1837–1865.

48. Wang, C., et al., *Interpretable machine learning for predicting heavy metal removal and optimizing biochar characteristics.* Journal of Water Process Engineering, 2024. **68**: p. 106484.

49. Hassan, R., Z. Behtouei, and A. Baghban, *Advanced machine learning for precise prediction of biochar's heavy metal sorption efficiency.* Journal of Hazardous Materials Advances, 2025. **18**: p. 100739.

50. Barkhordari, M.S. and C. Qi, *Integrating machine learning and reliability analysis: A novel approach to predicting heavy metal removal efficiency using biochar.* Ecotoxicology and Environmental Safety, 2025. **299**: p. 118381.

51. Almalawi, A., et al., *Modeling of Remora Optimization with Deep Learning Enabled Heavy Metal Sorption Efficiency Prediction onto Biochar.* Chemosphere, 2022. **303**: p. 135065.

52. Jaffari, Z.H., et al., *Transformer-based deep learning models for adsorption capacity prediction of heavy metal ions toward biochar-based adsorbents.* Journal of Hazardous Materials, 2024. **462**: p. 132773.

53. Wei, X., et al., *Machine learning insights in predicting heavy metals interaction with biochar.* Biochar, 2024. **6**(1): p. 10.

54. Zhao, C., et al., *Predictive modeling of heavy metal lead and cadmium adsorption on biochar based on machine learning.* International Journal of Phytoremediation, 2026. **28**(4): p. 748–755.

55. Li, X., et al. *Predictive Machine Learning Model to Assess the Adsorption Efficiency of Biochar-Heavy Metals for Effective Remediation of Soil–Plant Environment*. Toxics, 2024. **12**, 575 DOI: 10.3390/toxics12080575.

56. Tan, L., et al., *Adsorption performance of Ni(II) by KOH-modified biochar derived from different microalgae species.* Bioresource Technology, 2024. **394**: p. 130287.

57. Cui, C., et al., *A N-doped red mud-biochar magnetic composite for highly efficient Cd2+ removal: Adsorption performance and mechanism.* Journal of Analytical and Applied Pyrolysis, 2026. **195**: p. 107705.

58. Dad, F.P., et al., *Adsorption of trace heavy metals through organic compounds enriched biochar using isotherm adsorption and kinetic models.* Environmental Research, 2024. **241**: p. 117702.

59. Wang, H., et al., *Enhanced removal of Cr(VI) from aqueous solution by nano- zero-valent iron supported by KOH activated sludge-based biochar.* Colloids and Surfaces A: Physicochemical and Engineering Aspects, 2022. **651**: p. 129697.

60. Wu, J., et al., *Insight into mass transfer mechanism and equilibrium modeling of heavy metals adsorption on hierarchically porous biochar.* Separation and Purification Technology, 2022. **287**: p. 120558.

61. Flórez, E., C. Jimenez-Orozco, and N. Acelas, *Unravelling the influence of surface functional groups and surface charge on heavy metal adsorption onto carbonaceous materials: An in-depth DFT study.* Materials Today Communications, 2024. **39**: p. 108647.