



OPEN

# Machine learning-powered estimation of malachite green photocatalytic degradation with NML-BiFeO<sub>3</sub> composites

Iman Salahshoori<sup>1,2</sup>, Amirhosein Yazdanbakhsh<sup>3</sup> & Alireza Baghban<sup>4</sup>✉

This study explores the potential of photocatalytic degradation using novel NML-BiFeO<sub>3</sub> (noble metal-incorporated bismuth ferrite) compounds for eliminating malachite green (MG) dye from wastewater. The effectiveness of various Gaussian process regression (GPR) models in predicting MG degradation is investigated. Four GPR models (Matern, Exponential, Squared Exponential, and Rational Quadratic) were employed to analyze a dataset of 1200 observations encompassing various experimental conditions. The models have considered ten input variables, including catalyst properties, solution characteristics, and operational parameters. The Exponential kernel-based GPR model achieved the best performance, with a near-perfect R<sup>2</sup> value of 1.0, indicating exceptional accuracy in predicting MG degradation. Sensitivity analysis revealed process time as the most critical factor influencing MG degradation, followed by pore volume, catalyst loading, light intensity, catalyst type, pH, anion type, surface area, and humic acid concentration. This highlights the complex interplay between these factors in the degradation process. The reliability of the models was confirmed by outlier detection using William's plot, demonstrating a minimal number of outliers (66–71 data points depending on the model). This indicates the robustness of the data utilized for model development. This study suggests that NML-BiFeO<sub>3</sub> composites hold promise for wastewater treatment and that GPR models, particularly Matern-GPR, offer a powerful tool for predicting MG degradation. Identifying fundamental catalyst properties can expedite the application of NML-BiFeO<sub>3</sub>, leading to optimized wastewater treatment processes. Overall, this study provides valuable insights into using NML-BiFeO<sub>3</sub> compounds and machine learning for efficient MG removal from wastewater.

**Keywords** Dye removal, Kernel-based Gaussian process regression (GPR), Metal-incorporated bismuth ferrite (BiFeO<sub>3</sub>), Machine learning, Photocatalytic degradation, Wastewater treatment

Water pollution, the introduction of harmful substances into water bodies like rivers, lakes, and oceans, stems from industrial processes, agriculture, urban runoff, and sewage disposal<sup>1–12</sup>. This pollution jeopardizes human health, ecosystems, economic activities, and access to clean drinking water<sup>13,14</sup>. Addressing water pollution is crucial for environmental justice, combating climate change, and sustaining a healthy future<sup>15–18</sup>. Pollutants such as organic compounds, pharmaceuticals, and chemicals harm aquatic life and water quality, highlighting the need for effective management and regulation to protect both the environment and human health<sup>19–24</sup>.

Traditional wastewater treatment can struggle with eliminating persistent organic pollutants due to their resistance to conventional methods, complex molecular structures, and the potential formation of harmful byproducts during treatment<sup>25–27</sup>. To overcome these limitations, advanced technologies like adsorption<sup>28,29</sup>, membrane filtration<sup>30</sup>, biological treatment<sup>31</sup>, and Advanced Oxidation Processes (AOPs)<sup>32,33</sup> are being developed and tailored to specific pollutant profiles with ongoing regulatory updates to protect the environment and public health<sup>34,35</sup>. Each method carries distinct benefits and drawbacks. Adsorption efficiently eliminates heavy metals, organic compounds, and dye with minimal maintenance; yet, it demands expensive adsorbent material replacement and lacks universality in pollutant removal. Membrane filtration is effective but entails high costs

<sup>1</sup>Department of Polymer Processing, Iran Polymer and Petrochemical Institute, PO Box 14965-115, Tehran, Iran. <sup>2</sup>Department of Chemical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran. <sup>3</sup>Department of Polymer Engineering, School of Chemical Engineering, College of Engineering, University of Tehran, Tehran, Iran. <sup>4</sup>Department of Process Engineering, NISOC Company, Ahvaz, Iran. ✉email: Alireza\_baghban@ut.ac.ir

due to maintenance and fouling concerns. Biological treatments, such as activated sludge and trickling filters, are effective but require ample space and meticulous handling. The appropriate system selection depends on factors like pollutant type, effluent quality, economic feasibility, and environmental repercussions<sup>36–41</sup>.

AOPs, notably photocatalysis, are recognized for effectively combating toxic organic pollutants in environmental and wastewater treatment<sup>42</sup>. Photocatalysis offers eco-friendly, efficient, and cost-effective solutions by utilizing natural energy sources to generate reactive species for sustainable water treatment and environmental remediation<sup>43</sup>.

Bismuth ferrite (BiFeO<sub>3</sub>), a magnetic perovskite, shows promise in photocatalysis due to its low bandgap energy, thermal and chemical stress resistance, non-toxicity, and visible light responsiveness<sup>44,45</sup>. However, rapid recombination of photogenerated charge carriers limits its practical use<sup>46</sup>. Addressing this challenge involves strategically incorporating noble metals like silver (Ag), platinum (Pt), and palladium (Pd) as co-catalysts on BiFeO<sub>3</sub>'s surface<sup>47</sup>. This prevents electron loss to noble metals, enhancing electron management and catalytic efficiency<sup>21,48</sup>. The accumulation of electrons at noble metal surfaces facilitates reduction reactions, while BiFeO<sub>3</sub>'s valence band holes generate reactive hydroxyl radicals, crucial for chemical transformations<sup>46,49</sup>. Composite materials of BiFeO<sub>3</sub> with noble metals efficiently degrade organic pollutants, outperforming BiFeO<sub>3</sub> alone<sup>46,48</sup>.

The efficiency of a photocatalyst in degrading pollutants is influenced by various factors, including the catalyst's properties (such as pore volume and surface area), pollutant characteristics (like composition and concentration), competing compounds, and reaction conditions (e.g., time, pH, catalyst dosage, and light intensity)<sup>50</sup>. Achieving optimal conditions through experimentation can take days to months, especially for degrading MG dyes. However, the standard empirical method may not capture complex interactions among factors affecting efficiency<sup>51,52</sup>. Many photocatalytic materials, especially those responsive to visible light like BiFeO<sub>3</sub>, demonstrate superior performance compared to TiO<sub>2</sub>, but are associated with high costs. Developing an analytical, data-centric template could enhance the optimization of the photocatalytic process, improving its economic feasibility<sup>53</sup>. Such an approach would consider the interconnectedness of factors influencing water quality, streamlining optimization efforts and contributing to the process's economic viability.

Machine learning (ML) strategically deploys mathematical algorithms to build predictive models from datasets, aiming to inform decisions across qualitative and quantitative dimensions<sup>54</sup>. Until now, the wastewater treatment domain has effectively implemented a variety of basic ML algorithms<sup>55,56</sup>. Gaussian Process Regression (GPR) offers a distribution-free approach, estimating values and uncertainty in predictions, which is ideal for the complex relationships encountered in wastewater treatment<sup>56</sup>.

Scientific studies have applied ML models to wastewater datasets to predict water quality, assess environmental impact, and evaluate treatment performance<sup>57–62</sup>. However, these studies often need more substantial evidence regarding the suitability of the chosen ML algorithm and utilize limited datasets, typically with few input variables. ML approaches may vary in function, leading to variability in estimate precision<sup>58,63</sup>. Therefore, selecting the right ML approach for predicting pollutant degradation in wastewater is crucial.

This study introduces a novel approach to predicting and comparing MG dye photodegradation productivity using a dataset of 1200 observations and ten input factors. It employs Gaussian Process Regression with four kernel functions and optimizes photocatalytic procedure settings across these factors. The research meticulously ensures the accuracy of ML models and analyzes the interdependencies among process factors for MG dye degradation. Post-processing techniques, including sensitivity analysis, evaluate feature effectiveness and shed light on individual input variables' significance in photodegradation. This multifaceted approach advances the understanding of wastewater treatment and provides a framework for future research in predictive modelling and process optimization.

## Computational methodology

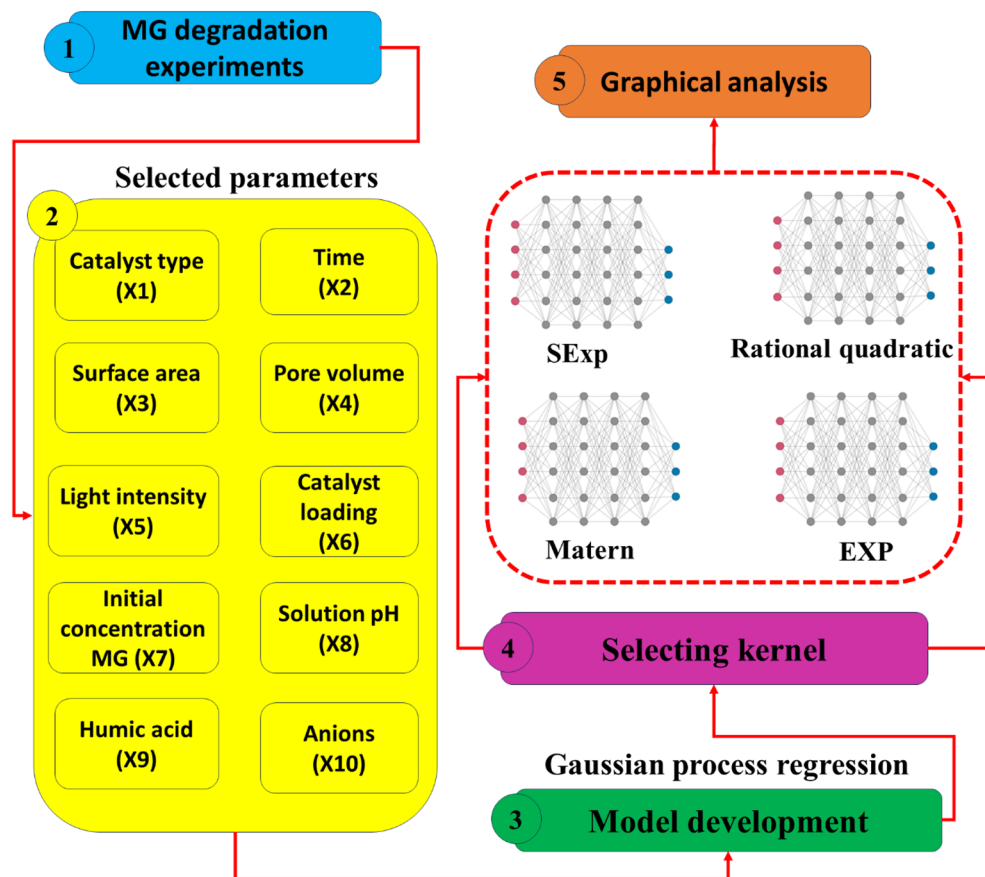
### Methodology

The methodology used in this study for modelling and optimizing MG dye photocatalysis using NML-BiFeO<sub>3</sub> compounds is depicted in Fig. 1. This methodology draws on insights from our previous research. Figure 1 illustrates that the study is conducted through three distinct stages. The initial step involves selecting ten parameters that significantly influence degradation efficiency, followed by designing and collecting 1200 data points through experimentation. In the second phase, an extensive comparison of four different kernel functions in the GPR model is conducted to identify the most suitable configurations for accurately predicting the efficiency of MG dye elimination throughout the photocatalytic procedure. Following this, the development of photocatalytic behaviour is determined by leveraging the four models exhibiting superior performance with higher R-squared values and lower error rates.

### Data preparation

We used 1200 data points in this investigation, which were obtained from a previous photocatalytic study<sup>64</sup>. Table S1 comprehensively lists the providers, concentrations, chemical formulas, labels, and intended applications of all chemicals used in this present investigation.

This study included a comprehensive set of 10 distinct features, meticulously selected because of their relevance and potential influence on the photocatalytic process under examination. These features encompassed a variety of parameters, including the type of catalyst employed, the duration of the experiment (in minutes), the surface area of the catalyst material (expressed in m<sup>2</sup>/g), the pore volume of the material (in cm<sup>3</sup>/g), the intensity of the illumination (in watts), the quantity of catalyst loaded into the system (in g/L), in-solution MG dye concentration (in mg/L), the pH of the solution, the concentration of humic acid (in mg/L), and the presence or absence of specific anions. The output variable was the efficacy of MG dye degradation.



**Figure 1.** Approach, gathering information, modelling, and additional processing schematic.

Within the data preparation phase, particular attention focused on two categorical input variables: the anions and catalyst types. We employed a new strategy to convert these attributes into numerical representations. To characterize catalyst types, a linear combination of the normalized surface area and pore volume of the catalysts was chosen. In addition, to characterize anion types, the normalized molecular weight of each anion was considered. It is worth noting that the normalization was carried out within the range of 0 to 1.

This conversion was deemed essential to ensure that the data met the stringent numerical prerequisites of ML algorithms, enabling seamless integration into the subsequent analytical processes. Preceding the commencement of machine learning model construction, a pivotal procedural step entailed randomly partitioning the dataset into two discrete subsets. Explicitly, 75% of the dataset was earmarked for utilization as the training dataset, whereas the remaining 25% was earmarked to serve as the test dataset. This division's rationale was to facilitate a comprehensive evaluation of the machine learning models post-training. This partitioning strategy ensured that the models were rigorously assessed on unseen data, gauging their generalization capabilities beyond the training phase.

### Gaussian process regression (GPR)

A powerful and well-structured machine learning approach, the GPR model, is well-regarded for its probabilistic and nonparametric characteristics. It can handle complex problems that involve non-linear relationships<sup>55</sup>. A key feature of this approach is the use of Gaussian processes for regression tasks. A significant aspect of its attractiveness arises from its capacity to efficiently incorporate uncertainty within its computational framework<sup>53</sup>.

In the context of GPR modelling, it is conventional to utilize two separate datasets: one allocated explicitly for training purposes (L) and another intended for testing (T). These datasets, T and L, are selected at random and comprise sets  $\{x_{L,i} \cdot y_{L,i}\}_{i=1}^n$ , and  $\{x_{T,i} \cdot y_{T,i}\}_{i=1}^n$ , where 'x' denotes the entered parameters and 'y' corresponds to the associated result factors. The following Equation establishes the basis of GPR modelling:

$$y_{L,i} = f(x_{L,i}) + \varepsilon_{L,i} \quad i = 1.2.3. \dots .n \quad (1)$$

$$\varepsilon \sim N(0 \cdot \sigma_{\text{noise}}^2 I_n) \quad (2)$$

Here, 'xL' signifies the individual factors, whereas 'yL' signifies the consequences linked to the training data sets. Furthermore, 'ε' serves as the notation for observation noise, 'σ<sup>2</sup><sub>noise</sub>' represents the noise variance, and 'I<sub>n</sub>' denotes the unit array in this context. In the same vein, we can articulate the following for the test dataset:

$$y_{T,i} = f(x_{T,i}) + \varepsilon_{T,i} \quad i = 1.2.3. \dots n \quad (3)$$

The symbols retain their previously defined meanings, but in this case, they pertain to the test dataset. Consequently, the Gaussian noise model links each computed ‘y’ value to the corresponding ‘f(x)’ function under consideration. As postulated by the GPR paradigm, ‘f(x)’ assumes the role of a stochastic function, and its characterization is contingent upon the concurrent utilization of the mean function ‘m(x)’ and the covariance function ‘k(x, x’),’ regularly recognized as kernel functions.

$$f(x_{L,i}) \sim GP(m(x) \cdot k(x \cdot x')) \quad (4)$$

It is possible to find the mean function “m(x)” by using specified basis functions; nonetheless, it is commonly approximated as zero for simplification and computational convenience<sup>66</sup>.

$$f(x_{L,i}) \sim GP(0 \cdot k(x \cdot x')) \quad (5)$$

Merging Eqs. (1) and (5) allows us to determine the ‘y’ distribution.

$$y \sim N(0 \cdot k(x \cdot x') + \sigma_{noise}^2 I_n) \quad (6)$$

Concluding the previously mentioned criteria and variables, the following deductions can be made:

$$\begin{bmatrix} \vec{f}_L \\ \vec{f}_T \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x_L, x_L) & k(x_L, x_T) \\ k(x_T, x_L) & k(x_T, x_T) \end{bmatrix}\right) \quad (7)$$

$$\begin{bmatrix} \vec{\varepsilon}_L \\ \vec{\varepsilon}_T \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{noise}^2 I_n & 0 \\ 0 & \sigma_{noise}^2 I_n \end{bmatrix}\right) \quad (8)$$

Incorporating the most recent pair of equations, we can derive the subsequent Gaussian expression:

$$\begin{bmatrix} \vec{y}_L \\ \vec{y}_T \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x_L, x_L) + \sigma_{noise}^2 I_n & k(x_L, x_T) \\ k(x_T, x_L) & k(x_T, x_T) + \sigma_{noise}^2 I_n \end{bmatrix}\right) \quad (9)$$

By applying the Gaussian conditioning principle, we can acquire the distribution for the variable ‘y<sub>T</sub>’:

$$(y_T | y_L) \sim N(\mu_T, \Sigma_T) \quad (10)$$

$$\Sigma_T = k(x_T, x_T) = k(x_T, x_T) + \sigma_{noise}^2 I_n - k(x_T, x_L)(k(x_L, x_L) + \sigma_{noise}^2 I_n)^{-1} k(x_L, x_T) \quad (11)$$

$$\mu_T = m\left(\frac{\rightarrow}{y_T}\right) = k(x_T, x_L)(k(x_L, x_L) + \sigma_{noise}^2 I_n)^{-1} \vec{y}_L \quad (12)$$

In this scenario,  $\Sigma_T$  represents the covariance, while  $\mu_T$  signifies the mean value. A GPR model’s predictive power and resilience are influenced by kernel function with a non-singular symmetric template. Four options—Squared exponential, Exponential, Matern, and Rational quadratic—have been selected to identify the best-suited kernel function. Presented below are the selected kernel functions:

Rational quadratic kernel function:

$$k_{RQ}(x, x') = \sigma^2 \left(1 + \frac{x - x'}{2a\ell}\right)^{-a} \quad (13)$$

Matern kernel function:

$$k_M(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{x - x'}{\ell}\right)^\nu K_\nu\left(\sqrt{2\nu} \frac{x - x'}{\ell}\right) \quad (14)$$

Squared Exponential kernel function:

$$k_{SE}(x, x') = \sigma^2 \exp\left(-\frac{x - x'}{\ell^2}\right) \quad (15)$$

Exponential kernel function:

$$k_E(x, x') = \sigma^2 \exp\left(-\frac{x - x'}{\ell}\right) \quad (16)$$

Within this context, the parameters  $\ell$ ,  $\sigma^2$ ,  $\sigma$ , and  $\alpha$  correspond to length scale, variance, amplitude and scale mixture, respectively. Furthermore, the symbols  $\nu$ ,  $\Gamma$ , and  $K_\nu$  were employed to signify a positive parameter, gamma function, and modified Bessel function, respectively.

### Performance metrics

The performance of the established models depended on the data quality and input factors. To measure model performance, a set of statistical measures were employed: the coefficient of determination ( $R^2$ ), root-mean-square error (RMSE), and mean absolute error (MAE). The subsequent equations delineate these parameters:

$$MAE = \frac{[\sum_{i=1}^n |o_i - p_i|]}{n} \quad (17)$$

$$RMSE = \sqrt{\frac{[\sum_{i=1}^n (o_i - p_i)^2]}{n}} \quad (18)$$

$$R^2 = 1 - \frac{\sum (o_i - p_i)^2 (p_i - \bar{p})}{\sum (o_i - \bar{o})^2 \sum (p_i - \bar{p})^2} \quad (19)$$

Here, "n" signifies the total number of samples considered. "oi" represents the observed removal efficiencies, whereas "pi" stands for the calculated removal efficacies. Furthermore, "p" holds the significance of being the mean value derived from all anticipated effectiveness quantities.

## Results and discussion

### Model development and testing

This study employed MATLAB software version 2018 to develop GPR models for predicting MG dye photocatalytic degradation. Table 1 compares our findings with previous research on organic pollutant degradation. The GPR models developed here achieved superior R-squared values and lower MAE and RMSE values compared to a significant portion of the existing literature. High R-squared values indicate strong agreement between predicted and experimental degradation values, validating the effectiveness of the models.

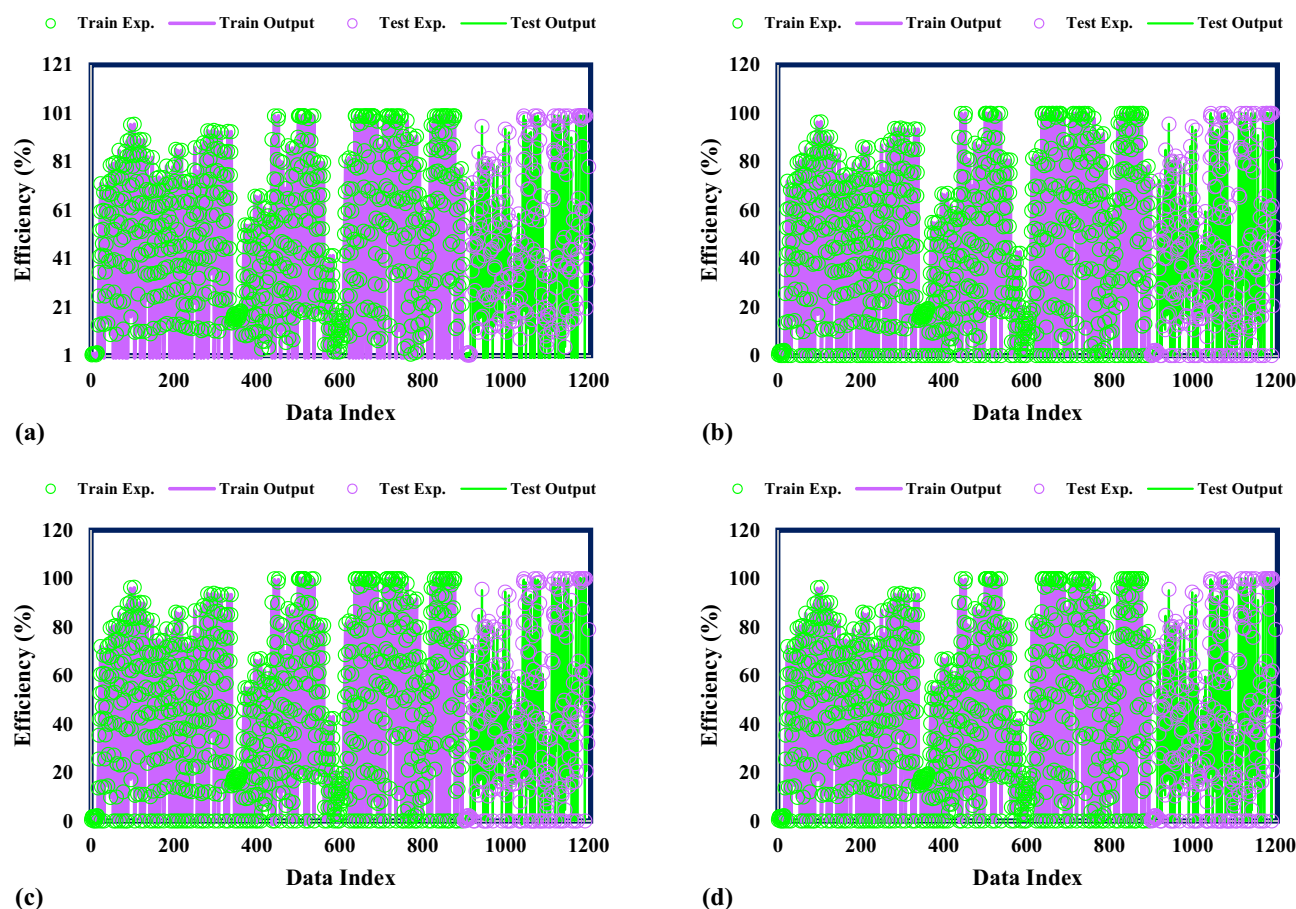
We examined error characteristics (STD, RMSE, MSE, MRE) to assess the training performance of the recommended GPR models. The error metrics indicate that the models effectively captured patterns and trends in the training data. Notably, the GPR model with an exponential kernel demonstrated excellent accuracy in predicting MG dye degradation for unseen data. Its high R-squared value (1.0) and low error metrics highlight its superior predictive capabilities.

This exceptional performance suggests the model's effectiveness in handling the complexities of MG dye photocatalytic degradation in wastewater, with potential applications in carbon capture and utilization. The multifaceted nature of the experimental design and the inclusion of diverse input features contribute to the richness and comprehensiveness of this study, leading to a more meaningful understanding of the underlying phenomena. Consequently, the GPR model's predictive performance emerges as a more reliable and suitable solution for addressing real-world challenges in this domain.

The correctness of the proven models is further validated by the simultaneous presentation of the anticipated and experimental values for the photocatalytic degradation of the MG dye in Fig. 2. Upon careful examination of the data, it is evident that the photocatalytic destruction of the experimental MG dye aligns with the many GPR models. This agreement precisely demonstrates the models' ability to predict the MG dye photocatalytic

Model	Group	R <sup>2</sup>	MRE (%)	MSE	RMSE	STD
Matern	Train data	1.000	0.005	4.83768E-06	0.0022	0.0016
	Test data	1.000	0.009	6.08233E-06	0.0025	0.0018
	Total data	1.000	0.006	5.14884E-06	0.0025	0.0016
Exponential	Train data	1.000	0.120	0.004360388	0.0660	0.0607
	Test data	1.000	0.230	0.00451456	0.0672	0.0624
	Total data	1.000	0.157	0.004398931	0.0672	0.0611
Squared exponential	Train data	1.000	0.363	0.039625283	0.1991	0.1667
	Test data	1.000	0.538	0.060361797	0.2457	0.2134
	Total data	1.000	0.435	0.044809411	0.2457	0.1795
Rational quadratic	Train data	1.000	1.413	0.269581839	0.5192	0.4295
	Test data	1.000	3.356	0.308675869	0.5556	0.4671
	Total data	1.000	2.007	0.279355346	0.5556	0.4390

**Table 1.** The numerical measures associated with the GPR models are indicated in this research.



**Figure 2.** Findings from experiments and the kernel-based GPR algorithm for (a) Matern, (b) Exponential, (c) Squared exponential, (d) Rational quadratic.

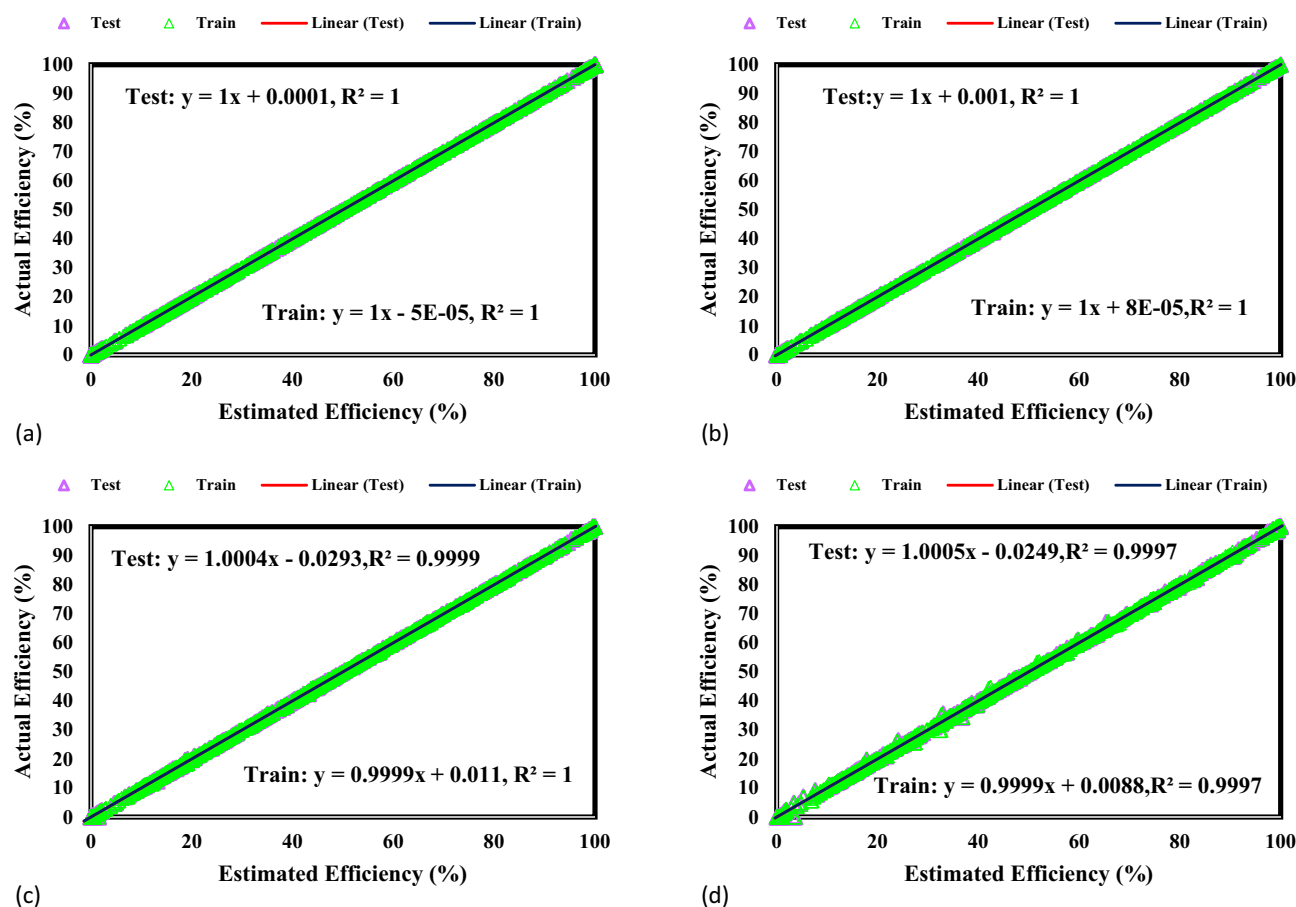
degradation in NML-BiFeO<sub>3</sub> composites. A broad investigation of the presented models shows a strong match between the anticipated and observed MG dye photocatalytic degradation rates.

This tight correlation shows that GPR models can accurately predict MG dye photocatalytic degradation in NML-BiFeO<sub>3</sub>. The algorithms' exact alignment between predicted and observed values shows their capacity to precisely capture photocatalytic degradation events, which could impact wastewater treatment. The remarkable effectiveness of GPR models enhances the field of model prediction as researchers gain more confidence in using these models to make predictions about MG dye removal efficiency and improve processes linked to photocatalytic degradation.

The visual representation in Fig. 3 illustrates the prediction accuracy of GPR models in the process of MG dye photocatalytic degradation compared to the data collected from experiments. The graph demonstrates an important link above 1,000 between the predicted and experimental outcomes.

The exact synchronization of the matching lines with the 45° line indicates the systems' accuracy in detecting complicated degradation trends. The precise positioning along the dividing line, especially in the GPR model using the Matern kernel function, achieves an impeccable correlation value of 1. The graph is an essential tool for evaluating the accuracy of GPR models in forecasting the photocatalytic degradation of MG dye within the NML-BiFeO<sub>3</sub> composite. Researchers gain vital knowledge on the accuracy of models, which helps improve wastewater treatment technologies and informs choices in academic and commercial contexts. The excellent accuracy shown by the Matern kernel-equipped GPR model distinguishes it as a noteworthy instrument for forecasting MG dye photocatalytic degradation with unprecedented precision.

Figure 4 illustrates and communicates crucial information about the predictive efficacy of GPR models in the context of MG dye photocatalytic degradation. The figure prominently displays the differences between experimentally measured MG dye photocatalytic degradation values and the corresponding estimated values obtained from GPR models. The accuracy of different GPR models is evaluated based on their ability to predict MG dye photocatalytic degradation. The Rational Quadratic and Squared Exponential kernel functions are highlighted for their remarkable accuracy. The relative deviation points for these models are reported to be below 30%, demonstrating a tight correlation across expected and investigational results. The relative deviation points of the Exponential kernel function are less than 1%, while the Matern kernel function stands out for its superior accuracy, showcasing absolute deviation points below 0.1%. This suggests a high precision in capturing the underlying behavior of photocatalytic degradation. The accuracy and reliability of the GPR models, especially those using specific kernel functions, are emphasized. This supports their credibility for predicting MG



**Figure 3.** Graphs depicting the Kernel-based GPR system for (a) Matern, (b) Exponential, (c) Squared exponential, (d) Rational quadratic.

dye photocatalytic degradation in the NML-BiFeO<sub>3</sub> composite. The discoveries indicate that this information could help scholars choose the most appropriate GPR systems for different purposes, particularly in wastewater treatment and employment inquiry. The overall aim is to contribute to sustainable solutions by improving the understanding and prediction of dye pollutant emissions.

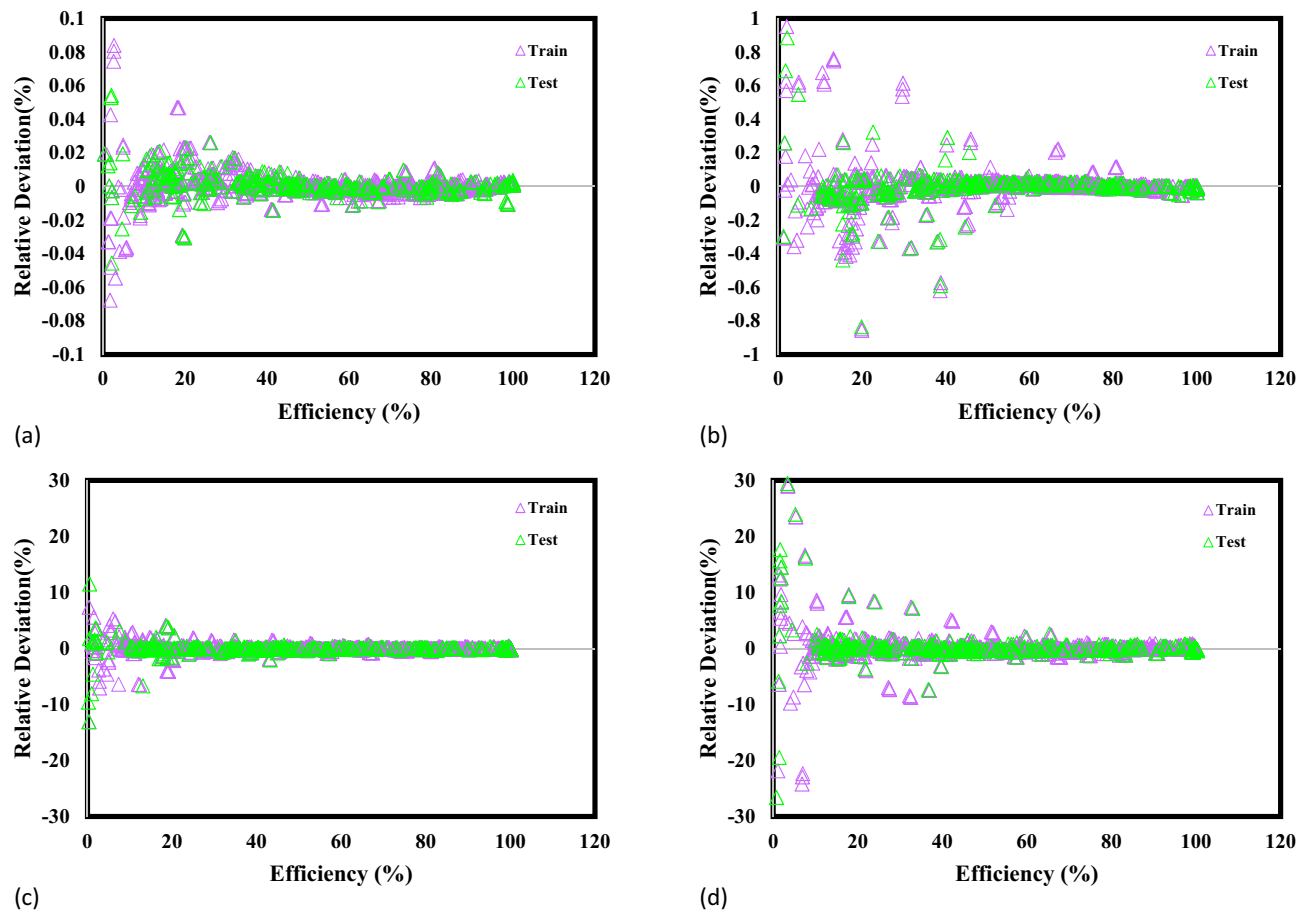
The insights from Fig. 4 regarding GPR models' MG dye photocatalytic degradation predictions are significant. The emphasis on kernel functions and accuracy levels helps scientists select the best models for specific functions, boosting wastewater treatment and sustainable solutions research. Figure 5 compares the current four GPR models with the models developed by Jaffari et al.<sup>64</sup> to estimate the efficiency of MG photocatalytic degradation with NML-BiFeO<sub>3</sub> composites. As can be seen, the current models achieve higher accuracy compared to the literature models, evidenced by lower errors and higher R-squared values.

### Sensitivity analysis

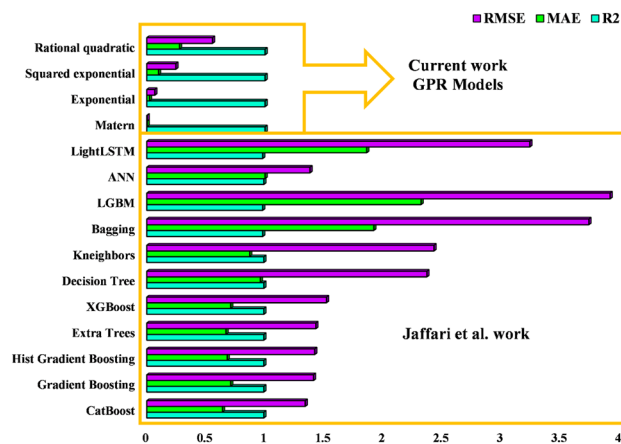
Sensitivity inquiry is conventionally carried out to explore the impact of input factors on the resultant output quantity<sup>67</sup>. As part of this in-depth analysis, it is imperative to consider the relevance factor, represented as 'r', which serves as the primary indicator of the input parameter exerting the most significant impact on MG photocatalytic degradation with NML-BiFeO<sub>3</sub> composites. This influential parameter can be quantified using the ensuing Equation:

$$r = \frac{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_{k,i} - \bar{X}_k)^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (20)$$

Within the presented framework, a variety of notations are employed, each possessing specific meanings:  $X_{k,i}$  is indicative of the 'kth' input parameter,  $\bar{X}_k$  represents the average value of input parameters,  $Y_i$  signifies the 'ith' output,  $\bar{Y}$  denotes the average of outputs, and 'n' denotes the total quantity of data points included in the analysis. Typically, the 'r' value exhibits variation within the range of -1 to +1. It is worth emphasizing that the absolute value of 'r' measures how each input variable affects the output variable. A higher absolute value of 'r' signifies a more pronounced correlation between each input and its output. Notably, negative values represent a situation where higher input values correspond to lower output values, while positive values indicate that higher



**Figure 4.** A comparison of the prediction performance of GPR models using (a) Exponential, (b) Matern, (c) Squared exponential, and (d) Rational quadratic versus empirical information.



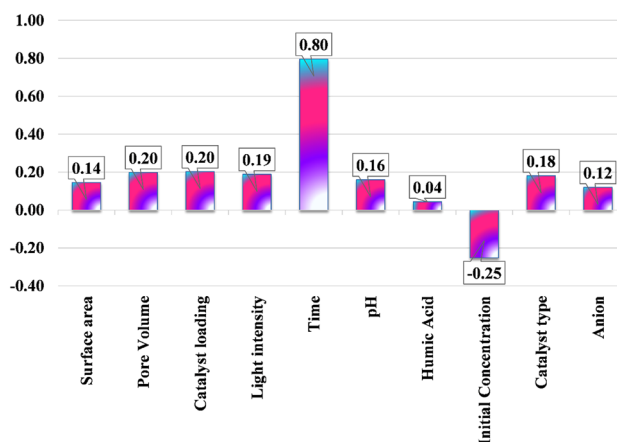
**Figure 5.** Statistical comparison of the current GPR models with the Jaffari et al.<sup>64</sup> models.

input values are associated with higher output values<sup>68</sup>. The work includes a visually captivating representation in Fig. 6, which is significant.

The sensitivity study illuminates the complex interplay between input parameters and MG photocatalytic degradation, successfully identifying the crucial factors that contribute to the process.

Analyzing feature significance with the GPR model allows us to comprehend the impact of operational factors on the photodegradation estimate of MG dye. Our investigation focused on understanding how various input features influenced the GPR model’s overall accuracy. Figure 6 presents the resulting assessment of the relative importance of these input features.





**Figure 6.** MG photocatalytic degradation with NML-BiFeO<sub>3</sub> composites input parameter analysis.

Pore volume and catalyst loading contribute 20% each, followed by light intensity at 19%. Catalyst type contributes 18%, followed by the pH of the solution at 16%. Anion type contributes 12%, surface area contributes 14%, and humic acid concentration contributes 4%. The most important factor in this situation is the photocatalytic process's time.

Notably, the gap in relative significance between the most critical factor, represented by time, and the least significant factor, exemplified by humic acid concentration, exceeds 80%. It becomes apparent that the degradation of MG dye was markedly impacted by the input factors linked to the circumstances of the photocatalytic process, as illustrated in the inset of Fig. 6. Further scrutiny of the GPR model involved a thorough investigation through a permutation significance assessment. This method discerns the decrement in model effectiveness resulting from the random reshuffling of an individual feature<sup>69</sup>. This procedure creates a disconnect between the input attributes and the effectiveness of MG dye degradation, leading inexorably to a downturn in the model's performance rating, thereby underscoring the model's dependence on these precise attributes.

### Outlier detection

Data points deemed outliers or giving rise to suspicion demonstrate dissimilar behaviour in comparison to the remaining data, and this disparity is frequently attributed to experimental irregularities or instrumental inaccuracies. To enhance the efficiency of the determined model and prevent erroneous analysis, it is imperative to identify and address potentially problematic data within the dataset. To streamline this procedure, we employ the Leverage method, a technique in which the Hat matrix is precisely articulated as follows:

$$H = U(U^T U)^{-1} U^T \quad (21)$$

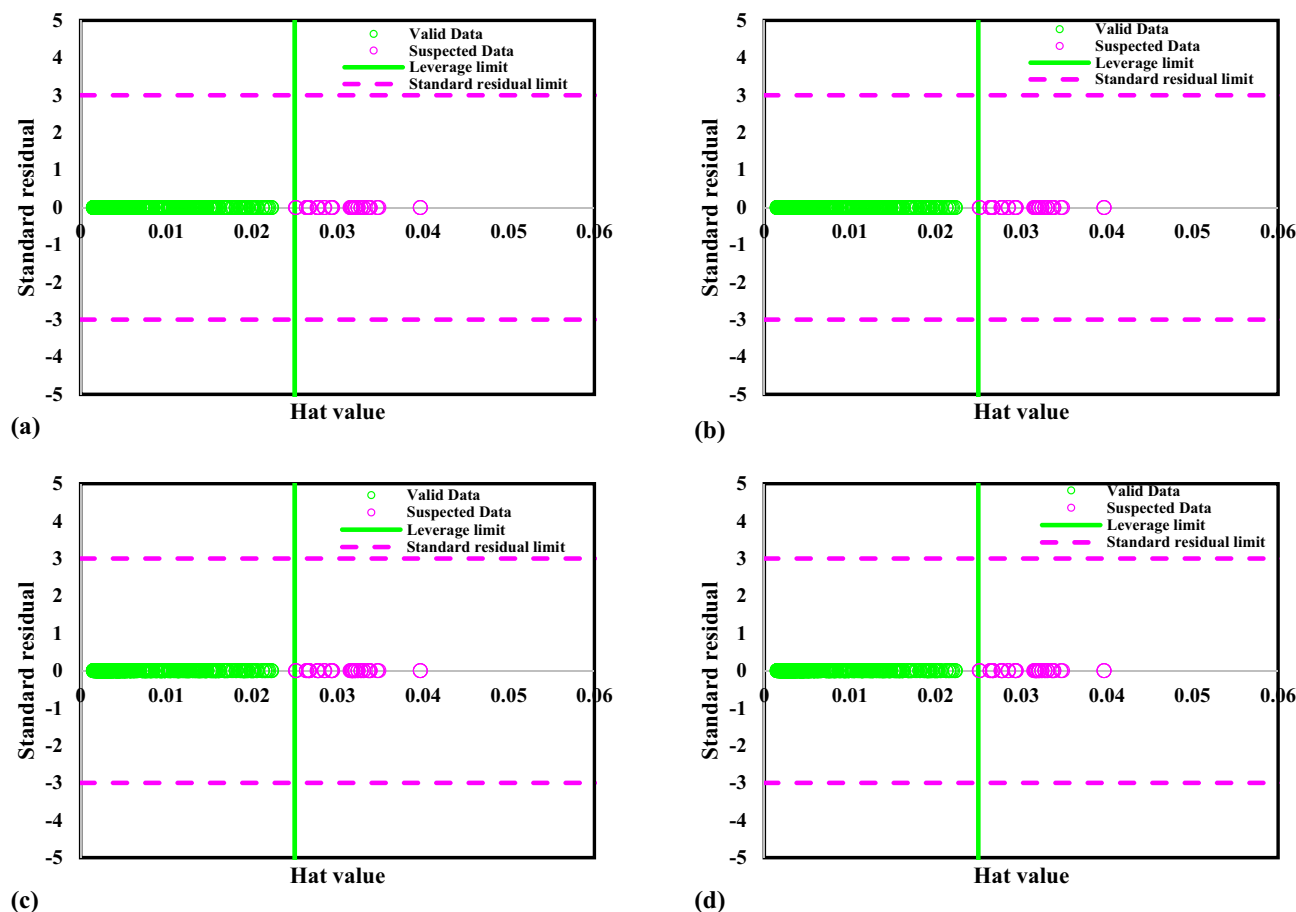
$U$  is characterized as a matrix with sizes  $i \times j$ , where  $i$  denotes the parameter count, and  $j$  represents the number of training data points. A visual depiction known as a Williams plot is produced to evaluate the veracity of the information. This analysis involves plotting standardized residuals against Hat values, allowing for a comprehensive evaluation; any data falling outside a designated region is considered potentially questionable. This dependable zone is a narrow space encompassing Hat values and residuals with a standard deviation between  $-3$  and  $3$ , ranging from 0 to the limits of critical leverage. The calculation for the limits of critical leverage is determined as follows<sup>70,71</sup>:

$$H^* = \frac{3(j+1)}{i} \quad (22)$$

Drawing insights from William's plot of the MG photocatalytic degradation data bank (Fig. 7), one can infer that a significant portion of the data employed in the analysis is deemed reliable. To provide a more detailed breakdown, out of a total of 1200 data points, only 71, 68, 69, and 66 outliers were identified for the GPR-Rational quadratic models, GPR-Squared Exponential, GPR-Exponential, and GPR-Matern, respectively.

### Implications and drawbacks of the current study

The utilization of NM-BiFeO<sub>3</sub> composites reveals considerable promise as a viable option for catalyzing the degradation of organic contaminants in aqueous environments. Experimental measurements involving controlled variables are usually employed in the conventional approach to establish the correlation between degradation effectiveness and reaction settings. However, these hands-on experiments often come with high costs, consume significant time, and need help achieving broad approval. This study employed four proficient ML models to illustrate the performance of MG dye photodegradation. This highlights a notable potential for promptly forecasting empirical outcomes using predetermined settings. The study also identified the key attributes of a photocatalyst's surface characteristics. It assessed their influence on the material's effectiveness in degrading organic pollutants and facilitating selective conditions during photocatalytic reactions for treating



**Figure 7.** William's MG photocatalytic degradation data bank visualization for outliers for Kernel-based GPR model of (a) Matern, (b) Exponential, (c) Squared exponential, (d) Rational quadratic.

organic wastewater. Applying this method will substantially diminish the necessity for extensive experimental exploration, resulting in cost savings and expediting the utilization of NML-BiFeO<sub>3</sub> compounds in organic wastewater treatment.

The current investigation underscores ML as a promising avenue for forecasting NML-BiFeO<sub>3</sub>-assisted photodegradation of MG dye compounds under controlled parameters. However, it is important to acknowledge limitations. Photocatalytic performance can be significantly influenced by various other factors, including temperature, pore volume, and catalyst loading. Additionally, this study does not account for the presence of multiple organic contaminants within a real-world wastewater treatment scenario. Fluctuations in these parameters could introduce discrepancies in the model, modify the significance of features, and limit the model's generalizability due to the absence of experimental data for these conditions. Future research will prioritize understanding the influence of these variables on the NML-BiFeO<sub>3</sub> photocatalytic process. The model will be further refined by incorporating additional data to enhance its precision and broaden its applicability to a wider range of organic pollutants. It is important to note that different organic pollutants may behave differently within photocatalytic systems. Therefore, further exploration using readily available datasets and a comprehensive investigation of these variables' influence on the photocatalytic breakdown of various organic pollutants in wastewater is warranted.

## Conclusions

In this study, we investigated the potential of various Gaussian process regression (GPR) models for predicting malachite green (MG) dye degradation using noble metal-incorporated bismuth ferrite (BiFeO<sub>3</sub>) (NML-BiFeO<sub>3</sub>) photocatalysts. The GPR models significantly outperformed existing methods in predicting MG degradation efficacy, achieving exceptional accuracy. This high accuracy is validated by the high R<sup>2</sup> values and low error metrics. The exponential kernel-based GPR model demonstrated the most exceptional performance, with a near-perfect R<sup>2</sup> value of 1.0 and minimal errors. This establishes its exceptional suitability for forecasting MG photocatalytic degradation in wastewater treatment. The close alignment between predicted and experimental results underscores the reliability of the GPR models in estimating degradation rates. This precision strengthens the foundation for utilizing GPR models to guide decision-making and optimize processes related to MG dye degradation.

Notably, the Rational Quadratic and Squared Exponential kernel models exhibited significant accuracy, with deviations below 30%. The Exponential kernel achieved exceptional precision with less than 1% deviation, while

the Matern kernel surpassed all others with a deviation of less than 0.1%. These findings highlight the remarkable accuracy of these models, particularly those employing specific kernels, for predicting MG dye degradation using NML-BiFeO<sub>3</sub> photocatalysts. These insights empower researchers to select the most appropriate GPR systems for wastewater treatment applications, ultimately contributing to advancements in sustainability efforts.

Furthermore, the study identified crucial input factors influencing MG photocatalytic degradation through a comprehensive sensitivity analysis. The direct correlation between the input parameters and the degradation process reveals the complex interplay between these factors. Analyzing feature significance using the GPR model revealed that process time is the most influential factor, followed by pore volume, catalyst loading, light intensity, catalyst type, pH, anion type, surface area, and humic acid concentration.

The reliability of the data employed in the analysis is further supported by insights gleaned from William's plot. Notably, a minimal portion of the 1200 data points (ranging from 66 to 71 data points depending on the GPR model) were identified as outliers. This signifies the robustness of the data employed for model development.

In conclusion, this study demonstrates the promising potential of NML-BiFeO<sub>3</sub> composites for catalyzing the degradation of organic contaminants in wastewater. The utilization of GPR models for forecasting MG dye photodegradation offers a powerful tool for rapid and efficient prediction of empirical outcomes. Identifying key catalyst surface properties can significantly expedite the application of NML-BiFeO<sub>3</sub> in organic wastewater treatment, leading to reduced costs and streamlined experimental procedures. Future research endeavors should explore the incorporation of additional variables to further enhance model accuracy and broaden applicability.

## Data availability

All the data used for model development provides in supplemental information.

Received: 14 February 2024; Accepted: 5 April 2024

Published online: 15 April 2024

## References

- Asghari, M. & Salahshoori, I. Iran's petrochemical plant affects wetlands. *Science* **381**(6663), 1164–1164. <https://doi.org/10.1126/science.adk2462> (2023).
- Purrostam, S. *et al.* Melamine functionalized mesoporous silica SBA-15 for separation of chromium (VI) from wastewater. *Mater. Chem. Phys.* **307**, 128240. <https://doi.org/10.1016/j.matchemphys.2023.128240> (2023).
- Lin, L., Yang, H. & Xu, X. Effects of water pollution on human health and disease heterogeneity: A review. *Front. Environ. Sci.* <https://doi.org/10.3389/fenvs.2022.880246> (2022).
- Salahshoori, I. *et al.* MIL-53 (Al) nanostructure for non-steroidal anti-inflammatory drug adsorption in wastewater treatment: Molecular simulation and experimental insights. *Process Saf. Environ. Protect.* **175**, 473–494. <https://doi.org/10.1016/j.psep.2023.05.046> (2023).
- Montazeri, N. *et al.* pH-Sensitive adsorption of gastrointestinal drugs (famotidine and pantoprazole) as pharmaceutical pollutants by using the Au-doped@ZIF-90-glycerol adsorbent: insights from computational modeling. *J. Mater. Chem. A* **11**(47), 26127–26151. <https://doi.org/10.1039/D3TA05221D> (2023).
- Salahshoori, I. *et al.* Insights into the adsorption properties of mixed matrix membranes (Pebax 1657-g-Chitosan-PVDF-Bovine Serum Albumin@ZIF-CO<sub>3</sub>-1) for the Antiviral COVID-19 treatment drugs remdesivir and nirmatrelvir: An in silico study. *ACS Appl. Mater. Interfaces* **15**(26), 31185–31205. <https://doi.org/10.1021/acsami.3c03943> (2023).
- Yazdanbakhsh, A., Behzadi, A., Moghaddam, A., Salahshoori, I. & Khonakdar, H. A. Mechanisms and factors affecting the removal of minocycline from aqueous solutions using graphene-modified resorcinol formaldehyde aerogels. *Sci. Rep.* **13**(1), 22771. <https://doi.org/10.1038/s41598-023-50125-0> (2023).
- Javdani-Mallak, A. & Salahshoori, I. Environmental pollutants and exosomes: A new paradigm in environmental health and disease. *Sci. Total Environ.* **925**, 171774. <https://doi.org/10.1016/j.scitotenv.2024.171774> (2024).
- Salahshoori, I. & NamayandehJorabchi, M. Iran's Zayandeh Rud River basin in crisis. *Science* **382**(6677), 1369–1369. <https://doi.org/10.1126/science.adm8965> (2023).
- Salahshoori, I., Seyfaee, A., Babapoor, A. & Cacciotti, I. Recovery of manganese ions from aqueous solutions with cyanex 272 using emulsion liquid membrane technique: A design of experiment study. *J. Sustain. Metall.* **7**(3), 1074–1090. <https://doi.org/10.1007/s40831-021-00396-6> (2021).
- Asl, M. D., Iman Salahshoori, A. S., Ali Hatami, A. A. & Golbarari, .. Experimental results and optimization via design of experiment (DOE) of the copper ion recovery from aqueous solutions using emulsion liquid membrane (ELM) method. *Desalin. Water Treat.* **204**, 238–256. <https://doi.org/10.5004/dwt.2020.26280> (2020).
- Salahshoori, I., Hatami, A. & Seyfaee, A. Investigation of experimental results and D-optimal design of hafnium ion extraction from aqueous system using emulsion liquid membrane technique. *J. Iran. Chem. Soc.* **18**(1), 87–107. <https://doi.org/10.1007/s13738-020-02007-9> (2021).
- Salahshoori, I. *et al.* An in silico study of sustainable drug pollutants removal using carboxylic acid functionalized-MOF nanostructures (MIL-53 (Al)-(COOH)<sub>2</sub>): Towards a greener future. *Desalination* **559**, 116654. <https://doi.org/10.1016/j.desal.2023.116654> (2023).
- Mousavi, S. R., Asghari, M., Mahmoodi, N. M. & Salahshoori, I. Water decolorization and antifouling melioration of a novel PEBA1657/PES TFC membrane using chitosan-decorated graphene oxide fillers. *J. Environ. Chem. Eng.* **11**(3), 109955. <https://doi.org/10.1016/j.jece.2023.109955> (2023).
- Salahshoori, I. *et al.* Study of modified PVDF membranes with high-capacity adsorption features using Quantum mechanics, Monte Carlo, and Molecular Dynamics Simulations. *J. Mol. Liquids* **375**, 121286. <https://doi.org/10.1016/j.molliq.2023.121286> (2023).
- Morin-Crimi, N. *et al.* Worldwide cases of water pollution by emerging contaminants: a review. *Environ. Chem. Lett.* **20**(4), 2311–2338. <https://doi.org/10.1007/s10311-022-01447-4> (2022).
- Gupta, B., Gupta, A. K., Tiwary, C. S. & Ghosal, P. S. A multivariate modeling and experimental realization of photocatalytic system of engineered S-C<sub>3</sub>N<sub>4</sub>/ZnO hybrid for ciprofloxacin removal: Influencing factors and degradation pathways. *Environ. Res.* **196**, 110390. <https://doi.org/10.1016/j.envres.2020.110390> (2021).
- Gupta, B. & Gupta, A. K. Photocatalytic performance of 3D engineered chitosan hydrogels embedded with sulfur-doped C<sub>3</sub>N<sub>4</sub>/ZnO nanoparticles for Ciprofloxacin removal: Degradation and mechanistic pathways. *Int. J. Biol. Macromol.* **198**, 87–100. <https://doi.org/10.1016/j.ijbiomac.2021.12.120> (2022).
- Salahshoori, I. *et al.* Assessing cationic dye adsorption mechanisms on MIL-53 (Al) nanostructured MOF materials using quantum chemical and molecular simulations: Toward environmentally sustainable wastewater treatment. *J. Water Process Eng.* **55**, 104081. <https://doi.org/10.1016/j.jwpe.2023.104081> (2023).

20. Qadafi, M., Wulan, D. R., Notodarmojo, S. & Zevi, Y. Characteristics and treatment methods for peat water as clean water sources: A mini review. *Water. Cycle* **4**, 60–69. <https://doi.org/10.1016/j.watcyc.2023.02.005> (2023).
21. Chaker, H., Chérif-Aouali, L., Khaoulani, S., Bengueddach, A. & Fourmentin, S. Photocatalytic degradation of methyl orange and real wastewater by silver doped mesoporous TiO<sub>2</sub> catalysts. *J. Photochem. Photobiol. A: Chem.* **318**, 142–149. <https://doi.org/10.1016/j.jphotochem.2015.11.025> (2016).
22. Salahshoori, I. *et al.* Advancements in wastewater Treatment: A computational analysis of adsorption characteristics of cationic dyes pollutants on amide Functionalized-MOF nanostructure MIL-53 (Al) surfaces. *Sep. Purif. Technol.* **319**, 124081. <https://doi.org/10.1016/j.seppur.2023.124081> (2023).
23. Gupta, B., Gupta, A. K., Ghosal, P. S. & Tiwary, C. S. Photo-induced degradation of bio-toxic Ciprofloxacin using the porous 3D hybrid architecture of an atomically thin sulfur-doped g-C<sub>3</sub>N<sub>4</sub>/ZnO nanosheet. *Environ. Res.* **183**, 109154. <https://doi.org/10.1016/j.envres.2020.109154> (2020).
24. Gupta, B., Gupta, A. K. & Bhatnagar, A. Treatment of pharmaceutical wastewater using photocatalytic reactor and hybrid system integrated with biofilm based process: Mechanistic insights and degradation pathways. *J. Environ. Chem. Eng.* **11**(1), 109141. <https://doi.org/10.1016/j.jece.2022.109141> (2023).
25. Morin-Crini, N. *et al.* Removal of emerging contaminants from wastewater using advanced treatments. A review. *Environ. Chem. Lett.* **20**(2), 1333–1375. <https://doi.org/10.1007/s10311-021-01379-5> (2022).
26. Mukhopadhyay, A., Duttgupta, S. & Mukherjee, A. Emerging organic contaminants in global community drinking water sources and supply: A review of occurrence, processes and remediation. *J. Environ. Chem. Eng.* **10**(3), 107560. <https://doi.org/10.1016/j.jece.2022.107560> (2022).
27. Dutta, S., Gupta, B., Srivastava, S. K. & Gupta, A. K. Recent advances on the removal of dyes from wastewater using various adsorbents: a critical review. *Mater. Adv.* **2**(14), 4497–4531. <https://doi.org/10.1039/D1MA00354B> (2021).
28. Haciosmanoğlu, G. G. *et al.* Antibiotic adsorption by natural and modified clay minerals as designer adsorbents for wastewater treatment: A comprehensive review. *J. Environ. Manag.* **317**, 115397. <https://doi.org/10.1016/j.jenvman.2022.115397> (2022).
29. Wang, W. & Wang, A. Perspectives on green fabrication and sustainable utilization of adsorption materials for wastewater treatment. *Chem. Eng. Res. Des.* **187**, 541–548. <https://doi.org/10.1016/j.cherd.2022.09.006> (2022).
30. Cevallos-Mendoza, J., Amorim, C.G., Rodríguez-Díaz, J.M., Montenegro, M.d.C.B.S.M., Removal of contaminants from water by membrane filtration: A review, membranes (2022).
31. Zhou, Q., Sun, H., Jia, L., Wu, W. & Wang, J. Simultaneous biological removal of nitrogen and phosphorus from secondary effluent of wastewater treatment plants by advanced treatment: A review. *Chemosphere* **296**, 134054. <https://doi.org/10.1016/j.chemosphere.2022.134054> (2022).
32. Feijoo, S., Yu, X., Kamali, M., Appels, L. & Dewil, R. Generation of oxidative radicals by advanced oxidation processes (AOPs) in wastewater treatment: a mechanistic, environmental and economic review. *Rev. Environ. Sci. Bio/Technol.* **22**(1), 205–248. <https://doi.org/10.1007/s11157-023-09645-4> (2023).
33. Li, S. *et al.* Antibiotics degradation by advanced oxidation process (AOPs): Recent advances in ecotoxicity and antibiotic-resistance genes induction of degradation products. *Chemosphere* **311**, 136977. <https://doi.org/10.1016/j.chemosphere.2022.136977> (2023).
34. Bagherzadeh, F., Mehrani, M.-J., Basirifard, M. & Roostaei, J. Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance. *J. Water Process Eng.* **41**, 102033. <https://doi.org/10.1016/j.jwpe.2021.102033> (2021).
35. Rego, R. M., Kurkuri, M. D. & Kigga, M. A comprehensive review on water remediation using UiO-66 MOFs and their derivatives. *Chemosphere* **302**, 134845. <https://doi.org/10.1016/j.chemosphere.2022.134845> (2022).
36. Samy, M. *et al.* Treatment of hazardous landfill leachate containing 1,4 dioxane by biochar-based photocatalysts in a solar photo-oxidation reactor. *J. Environ. Manag.* **332**, 117402. <https://doi.org/10.1016/j.jenvman.2023.117402> (2023).
37. El-Bestawy, E. A., Gaber, M., Shokry, H. & Samy, M. Effective degradation of atrazine by spinach-derived biochar via persulfate activation system: Process optimization, mechanism, degradation pathway and application in real wastewater. *Environ. Res.* **229**, 115987. <https://doi.org/10.1016/j.envres.2023.115987> (2023).
38. Samy, M., Mensah, K., El-Fakharany, E. M., Elkady, M. & Shokry, H. Green valorization of end-of-life toner powder to iron oxide-nanographene nanohybrid as a recyclable persulfate activator for degrading emerging micropollutants. *Environ. Res.* **223**, 115460. <https://doi.org/10.1016/j.envres.2023.115460> (2023).
39. Mensah, K., Mahmoud, H., Fujii, M., Samy, M. & Shokry, H. Dye removal using novel adsorbents synthesized from plastic waste and eggshell: mechanism, isotherms, kinetics, thermodynamics, regeneration, and water matrices. *Biomass Convers. Biorefinery* <https://doi.org/10.1007/s13399-022-03304-4> (2022).
40. Salama, E. *et al.* The superior performance of silica gel supported nano zero-valent iron for simultaneous removal of Cr (VI). *Sci. Rep.* **12**(1), 22443. <https://doi.org/10.1038/s41598-022-26612-1> (2022).
41. Samy, M., Mensah, K. & Gar Alalm, M. A review on photodegradation mechanism of bio-resistant pollutants: Analytical methods, transformation products, and toxicity assessment. *J. Water Process Eng.* **49**, 103151. <https://doi.org/10.1016/j.jwpe.2022.103151> (2022).
42. Xia, H. *et al.* A review of microwave-assisted advanced oxidation processes for wastewater treatment. *Chemosphere* **287**, 131981. <https://doi.org/10.1016/j.chemosphere.2021.131981> (2022).
43. Chakraborty, J., Nath, I. & Verpoort, F. A physicochemical introspection of porous organic polymer photocatalysts for wastewater treatment. *Chem. Soc. Rev.* **51**(3), 1124–1138. <https://doi.org/10.1039/D1CS00916H> (2022).
44. Haruna, A., Abdulkadir, I. & Idris, S. O. Photocatalytic activity and doping effects of BiFeO<sub>3</sub> nanoparticles in model organic dyes. *Heliyon* **6**(1), e03237. <https://doi.org/10.1016/j.heliyon.2020.e03237> (2020).
45. Lam, S.-M., Sin, J.-C. & Mohamed, A. R. A newly emerging visible light-responsive BiFeO<sub>3</sub> perovskite for photocatalytic applications: A mini review. *Mater. Res. Bull.* **90**, 15–30. <https://doi.org/10.1016/j.materresbull.2016.12.052> (2017).
46. Zhang, Y., Cai, Z. & Ma, X. Photocatalysis enhancement of Au/BFO nanoparticles using plasmon resonance of Au NPs. *Physica B: Condensed Matter* **479**, 101–106. <https://doi.org/10.1016/j.physb.2015.09.045> (2015).
47. Lam, S.-M. *et al.* Insight into the influence of noble metal decorated on BiFeO<sub>3</sub> for 2,4-dichlorophenol and real herbicide wastewater treatment under visible light. *Colloids and Surf. A: Physicochem. Eng. Aspects* **614**, 126138. <https://doi.org/10.1016/j.colsurfa.2021.126138> (2021).
48. Niu, F. *et al.* Synthesis of Pt/BiFeO<sub>3</sub> heterostructured photocatalysts for highly efficient visible-light photocatalytic performances. *Solar Energy Mater. Solar Cells* **143**, 386–396. <https://doi.org/10.1016/j.solmat.2015.07.008> (2015).
49. Boudghene-Guerriche, A. *et al.* Evaluation of antibacterial and antioxidant activities of silver-decorated TiO<sub>2</sub> nanoparticles. *ChemistrySelect* **5**(36), 11078–11084. <https://doi.org/10.1002/slct.202002734> (2020).
50. Derikvandi, H. & Nezamzadeh-Ejehieh, A. Increased photocatalytic activity of NiO and ZnO in photodegradation of a model drug aqueous solution: Effect of coupling, supporting, particles size and calcination temperature. *J. Hazard. Mater.* **321**, 629–638. <https://doi.org/10.1016/j.jhazmat.2016.09.056> (2017).
51. Bassi, A., Hasan, I., Qanungo, K., Koo, B. H. & Khan, R. A. Visible light assisted mineralization of malachite green dye by green synthesized xanthan gum/agar@ZnO bionanocomposite. *J. Mol. Struct.* **1256**, 132518. <https://doi.org/10.1016/j.molstruc.2022.132518> (2022).

52. Sekar, A. & Yadav, R. Green fabrication of zinc oxide supported carbon dots for visible light-responsive photocatalytic decolourisation of Malachite Green dye: Optimization and kinetic studies. *Optik* **242**, 167311. <https://doi.org/10.1016/j.ijleo.2021.167311> (2021).
53. Zhu, X., Wang, X. & Ok, Y. S. The application of machine learning methods for prediction of metal sorption onto biochars. *J. Hazard. Mater.* **378**, 120727. <https://doi.org/10.1016/j.jhazmat.2019.06.004> (2019).
54. Hansen, L. D., Stokholm-Bjerregaard, M. & Durdevic, P. Modeling phosphorous dynamics in a wastewater treatment process using Bayesian optimized LSTM. *Comput. Chem. Eng.* **160**, 107738. <https://doi.org/10.1016/j.compchemeng.2022.107738> (2022).
55. Kim, M., Kim, Y., Kim, H., Piao, W. & Kim, C. Evaluation of the k-nearest neighbor method for forecasting the influent characteristics of wastewater treatment plant. *Front. Environ. Sci. Eng.* **10**(2), 299–310. <https://doi.org/10.1007/s11783-015-0825-7> (2016).
56. Wan, X. *et al.* Water quality prediction model using Gaussian process regression based on deep learning for carbon neutrality in papermaking wastewater treatment system. *Environ. Res.* **211**, 112942. <https://doi.org/10.1016/j.envres.2022.112942> (2022).
57. Firouzi, F. *et al.* Simultaneous adsorption-photocatalytic degradation of tetracycline by CdS/TiO<sub>2</sub> nanosheets/graphene nanocomposites: Experimental study and modeling. *J. Environ. Chem. Eng.* **9**(6), 106795. <https://doi.org/10.1016/j.jece.2021.106795> (2021).
58. Jiang, Z., Hu, J., Tong, M., Samia, A. C., Zhang, H., Yu, X. A novel machine learning model to predict the photo-degradation performance of different photocatalysts on a variety of water contaminants. *Catalysts* (2021).
59. Abdi, J., Hadipoor, M., Hadavimoghaddam, F. & Hemmati-Sarapardeh, A. Estimation of tetracycline antibiotic photodegradation from wastewater by heterogeneous metal-organic frameworks photocatalysts. *Chemosphere* **287**, 132135. <https://doi.org/10.1016/j.chemosphere.2021.132135> (2022).
60. Azadi, S., Karimi-Jashni, A. & Javadpour, S. Modeling and optimization of photocatalytic treatment of landfill leachate using tungsten-doped TiO<sub>2</sub> nano-photocatalysts: Application of artificial neural network and genetic algorithm. *Process Saf. Environ. Protect.* **117**, 267–277. <https://doi.org/10.1016/j.psep.2018.03.038> (2018).
61. Tabatabai-Yazdi, F.-S., EbrahimianPirbazari, A., Esmaili Khalil Saraei, F. & Gilani, N. Construction of graphene based photocatalysts for photocatalytic degradation of organic pollutant and modeling using artificial intelligence techniques. *Physica B: Condensed Matter* **608**, 412869. <https://doi.org/10.1016/j.physb.2021.412869> (2021).
62. Kassahun, S. K., Kiflie, Z., Kim, H. & Baye, A. F. Process optimization and kinetics analysis for photocatalytic degradation of emerging contaminant using N-doped TiO<sub>2</sub>-SiO<sub>2</sub> nanoparticle: Artificial Neural Network and Surface Response Methodology approach. *Environ. Technol. Innov.* **23**, 101761. <https://doi.org/10.1016/j.eti.2021.101761> (2021).
63. Gheytnazadeh, M. *et al.* An insight into tetracycline photocatalytic degradation by MOFs using the artificial intelligence technique. *Sci. Rep.* **12**(1), 6615. <https://doi.org/10.1038/s41598-022-10563-8> (2022).
64. Jaffari, Z. H. *et al.* Machine learning approaches to predict the photocatalytic performance of bismuth ferrite-based materials in the removal of malachite green. *J. Hazard. Mater.* **442**, 130031. <https://doi.org/10.1016/j.jhazmat.2022.130031> (2023).
65. Hoang, N.-D., Pham, A.-D., Nguyen, Q.-L. & Pham, Q.-N. Estimating compressive strength of high performance concrete with gaussian process regression model. *Adv. Civil Eng.* **2016**, 2861380. <https://doi.org/10.1155/2016/2861380> (2016).
66. Fu, Q. *et al.* Prediction of the diet nutrients digestibility of dairy cows using Gaussian process regression. *Inf. Proc. Agricult.* **6**(3), 396–406 (2019).
67. Gheytnazadeh, M. *et al.* Towards estimation of CO<sub>2</sub> adsorption on highly porous MOF-based adsorbents using gaussian process regression approach. *Sci. Rep.* **11**(1), 15710. <https://doi.org/10.1038/s41598-021-95246-6> (2021).
68. Baghban, A., Mohammadi, A. H. & Taleghani, M. S. Rigorous modeling of CO<sub>2</sub> equilibrium absorption in ionic liquids. *Int. J. Greenhouse Gas Control* **58**, 19–41. <https://doi.org/10.1016/j.ijggc.2016.12.009> (2017).
69. Ferreño, D., Serrano, M., Kirk, M., Sainz-Aja, J. A., Prediction of the transition-temperature shift using machine learning algorithms and the plotter database. *Metals* (2022).
70. Zhou, X., Zhou, F. & Naseri, M. An insight into the estimation of frost thermal conductivity on parallel surface channels using kernel based GPR strategy. *Sci. Rep.* **11**(1), 7203. <https://doi.org/10.1038/s41598-021-86607-2> (2021).
71. Razavi, R., Bemani, A., Baghban, A., Mohammadi, A. H. & Habibzadeh, S. An insight into the estimation of fatty acid methyl ester based biodiesel properties using a LSSVM model. *Fuel* **243**, 133–141. <https://doi.org/10.1016/j.fuel.2019.01.077> (2019).

## Author contributions

All authors contributed in software, analysis, data gathering, writing, and conception.

## Competing interests

The authors declare no competing interests.


## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-58976-x>.

**Correspondence** and requests for materials should be addressed to A.B.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024