Study of machine learning on the photocatalytic activity of a novel nanozeolite for the application in the Rhodamine B dye degradation

Leandro R. Oviedo1\*, Daniel M. Druzian1, Lissandro D. D. Nora1, William L. da Silva1

1Franciscan University, Applied Nanomaterials Research Group (GPNAp), Campus II, Building #11, Room 100C, ZIP: 97010492, Santa Maria - RS, Brazil

\**Corresponding author. leandro.oviedo@ufn.edu.br*

Abstract/Resumo

ABSTRACT – Contamination of wastewater with organic dyes has caused a serious threat to humans and aquatic life due to the hazardous effect of these contaminants. In this context, the present work aims to carry out a Machine Learning (ML) study to evaluate the photocatalytic activity of a nanozolite (nANA) in the degradation of Rhodamine B (RhB) dye. Three machine learning algorithms (Random Forest, Artificial Neural Network and Xtreme Gradient Boosting) were used in the regression model development. The dataset used in the machine learning and data correlation was generated by Central Composite Rotational Design (CCRD 2²). Regarding the machine learning study, the ANN with structure 3:6:1 showed the best performance as a predictive model (R² = 0.98 and 0.9 for training and testing, RMSE < 5.0), resulting in the 50.37 ± 1.01% RhB removal at pH 5.7, [RhB] = 200 mg L-1 and [nANA] = 2.75 g L-1 after 180 min under visible light. Feature importance revealed that all parameters (pH, [RhB], [nANA]) were relevant to the response. Therefore, this work confirms the potentiality of machine learning algorithms to develop predictive models as well as a good starting point for the scale-up of advanced oxidation processes.

*Keywords: nanozeolite, Rhodamine B, Machine Learning, heterogeneous photocatalysis, nanocatalysts.*

RESUMO - A contaminação de águas residuais com corantes orgânicos tem causado uma séria ameaça aos seres humanos e à vida aquática devido ao efeito perigoso desses contaminantes. Neste contexto, o presente trabalho tem como objetivo realizar um estudo de Machine Learning (ML) para avaliar a atividade fotocatalítica de uma nanozeólita (nANA) na degradação do corante Rodamina B (RhB). Três algoritmos de aprendizado de máquina (Random Forest, Artificial Neural Network e Xtreme Gradient Boosting) foram usados no desenvolvimento do modelo de regressão. O conjunto de dados usado no aprendizado de máquina e na correlação de dados foi gerado pelo delineamento composto central rotacional (DCCR 2²). Com relação ao estudo de aprendizado de máquina, a RNA com estrutura 3:6:1 apresentou o melhor desempenho como modelo preditivo (R² = 0,98 e 0,9 para treinamento e teste, RMSE < 5,0), resultando na remoção de 50,37 ± 1,01% de RhB em pH 5,7, [RhB] = 200 mg L-1 e [nANA] = 2,75 g L-1 após 180 min sob luz visível. A importância da característica revelou que todos os parâmetros (pH, [RhB], [nANA]) foram relevantes para a resposta. Por conseguinte, este trabalho confirma a potencialidade dos algoritmos de aprendizado de máquina para desenvolver modelos preditivos, bem como um bom ponto de partida para o aumento de escala dos processos oxidativos avançados.

*Palavras-chave: nanozeólita, Rodamina B, aprendizado de máquina, fotocatálise heterogênea, nanocatalisadores.*

## Introduction

The contamination of wastewater with organic pollutants has been considered a great concern to public organs, mainly due to the hazardous effects that they can cause to humans and aquatic life (1). For example, Rhodamine B (RhB) is a synthetic organic dye extensively used in textile industries and in the dyeing of semiprecious stones, such as agates in the south of Brazil (2). Moreover, RhB is characterized by its high water solubility, low biodegradability, toxicity and resistance to conventional physio-chemical and biological processes of wastewater treatments (3).

Thus, advanced treatment processes are necessary for the proper management of these persistent pollutants, with heterogeneous photocatalysis, due to the simplicity of operation, versatility and high efficiency in the removal of organic pollutants from aqueous solutions (4), through the generation of reactive species of oxygen (ROS), mainly the hydroxyl radical (●OH), responsible for a series of redox reactions on the surface of a catalyst under radiation (UV or visible) (5). Alternative nanocatalysts have been attractive due to their special properties (e.g., high surface area and porosity) (6). For example, low silica nanozeolites (analcime, sodalite, nanozeolite P) have been used for the cationic dye photodegradation (7). Moreover, these nanomaterials can be easily synthesized by hydrothermal method using alternative raw materials (such as alum sludge, rice husk and coal fly ash) (8). However, experimental procedures can be time and cost-consuming, requiring the use of alternative tools, such as computational tools.

Some machine learning algorithms, such as Xtreme Gradient Boosting (XGB), Random Forest (RF) and Artificial Neural Networks (ANN), have been used in some works associated with photocatalytic degradation (9). These algorithms can be successfully used to develop a generalized predictive model with high accuracy and low errors associated with the experimental procedures (10). The high computational power of these models is extremely useful in the predictions and data correlation, being suitable to prevent overfitting (11). In this context, the present work aims to develop a generalized predictive machine learning-based model with application in dye removal from aqueous solution by heterogeneous photocatalysis using an alternative low-silica nanozeolite as nanocatalyst for RhB degradation.

## Experimental section

*Synthesis and characterization of the nanocatalyst*

The nanostructured photocatalyst was synthesized by hydrothermal method, according to the literature (12). Thus, rice husk (RH) and alum sludge (AS) were used as precursors of the analcime nanozeolite (nANA). Initially, the residues were dissolved in NaOH (2 mol L-1, Synth) and placed in an autoclave stainless steel coated with polytetrafluoroethylene (PTFE) (at 180°C for 12 h). After, the nANA was filtered (ϕ = 0.45 µm) and washed with distilled water and ethanol 70% (until pH ≈ 7.00). The nANA was dried in an oven at 80°C for 12 h. More details about the characterization can be found in the literature (13).

*Photocatalytic assay*

The heterogeneous photocatalysis was carried out in a slurry reactor operated at 25 ± 2 °C under visible light (lamp power = 202 W m-2). The volume used was 50 mL in all experimental runs, which were obtained by central composite rotational design (CCRD) 22, according to Table 1. The pH was maintained in the natural pH of the aqueous solution containing RhB organic dye. All experimental runs were performed in duplicate.

**Table 1.** Dataset generated from CCRD 2²

|  |  |
| --- | --- |
| **Parameter** | **Range** |
| [RhB] (mg L-1) | 12; 20; 40; 60; 68 |
| [nANA] (g L-1) | 0.5; 1.0; 1.75, 2.75; 3.00 |

 pH range: 5.81 | T = 25 ± 2 °C.

*Kinetic of degradation*

The experimental runs were adjusted to the Langmuir-Hinshelwood model, according to the Eq. (1) (14):

|  |  |
| --- | --- |
|  | (1) |

Where *Ci0* and *Ci* are the initial concentration and concentration at time t, in mg L-1; *k* is the apparent rate of the pseudo first-order reaction (min-1); and *t* is reaction time (min).

*Machine learning*

To develop a generalized predictive model, three machine learning algorithms were used: Random Forest (RF), Artificial Neural Network (ANN) and Xtreme Gradient Boosting (XGB). The configurations tested in each model are informed in Table 2.

**Table 2.** Machine learning model configurations.

|  |  |  |
| --- | --- | --- |
| **Model** | **Configuration** | **Ref.** |
| RF | Maximum depth trees: 1, 2, 3, 5, 7, 10, 15 mNumber of decision trees: 25, 50, 75, 100, 150 | (15) |
| ANN | *3:n:1*, where *n* is a hidden layers of nodes ranging from 1-12.Activation functions: tanh and logisticWeights optimization: lbfgs\* and Adam\*\* function with back-propagation algorithm |
| XGB | Maximum depth trees: 1, 2, 3, 5, 7, 10, 15 mSampling: bootstrapping |

*\*Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm*

*\*\*Adaptive Movement Estimation*

Despite the variations in model parameters described in Table 2, a 5-fold cross validation (cv = 5) was used for all of them.

*Feature importance*

To evaluate the contribution of each parameter of the heterogeneous photocatalysis on the RhB degradation, a feature importance study measured by permutation score value was carried out (16). All permutation score above 0.15 was considered to cause great changes in the response of the machine learning predictive model, Removal (%).

*Inferential statistics*

To investigate the data distribution for each parameter used in the machine learning model, the Shapiro-Wilk normality test was carried out. It is useful to define the correlation method to be used in the data correlation (i.e., the relationship between the input variables with the output variables). Thus, assuming a normal distribution, a Pearson correlation method is used. Otherwise, a Spearman correlation method can be used (17).

*Data correlation*

To determine the correlation among the parameters of the photocatalytic degradation of RhB, a spearman (generally due to the non-normality of datasets used in the machine learning studies) correlation method was used. Thus, the negative and positive influences of each parameter on the removal percentage of RhB and *k*were identified.

## Results and Discussion

*Machine learning model selection*

Table 3 informs the performance of each machine learning algorithm used this work.

**Table 3.** Metrics evaluated for the machine learning models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training R2** | **Training RMSE** | **Testing R2** | **Testing RMSE** |
| RF | 0.8175 | 1.6131 | 0.2852 | 2.2886 |
| XGB | 0.9877 | 0.7436 | 0.1437 | 2.4550 |
| ANN | 0.9888 | 1.0786 | 0.90165 | 3.5524 |

According to Table 3, the ANN showed the best performance in the prediction of RhB removal under visible light, once the R2 for training and testing were similar. Moreover, the RMSE in both cases was low (close to 1). Additionally, testing R2 values greater than 0,80 indicates a very low probability of overfitting of the proposed ANN model (18). Furthermore, the low accuracy of RF and XGB (R2 < 30) was probably due to the dataset size used in these models, which is a small dataset (ca. 300 points) (19). Figure 1 shows the performance of the ANN model in the terms of number of neurons (nodes) used in the neural network structure.



**Figure 1.** ANN performance: correlation between number of nodes of the hidden layer and RMSE.

As can be seen in Figure 1, the error associated with ANN model reduces with the increase of the number of nodes in the hidden layer, showing an expressively reduction after 4 nodes. Moreover, the framework with nodes ranging from 6-12 nodes in the hidden layer showed no change in the RMSE. Thus, owing to lesser computational power, the neural network with 6 nodes in the hidden layer was used as the generalized predictive model, as illustrated in the diagram shown in Figure 2.



**Figure 2.** ANN structure generated for the prediction model.

According to Figure 2, the best ANN model configuration was 3:6:1, consisting of a network with 3 input parameters ([RhB], [nANA], pH), one hidden layer with 6 nodes (neurons) and 1 response variable (RhB Removal). The activation function that resulted in the best model performance was the hyperbolic tangent function, with the solver lbfgs function as the optimizer of weights used in the neural network. It was due to the size of the data set, which can be considered small. Then, the lbfgs is an optimizer function of the quasi-Newton family, which often converge faster and with better performance than the Adam function to relatively small datasets (ca. 300-400 datapoints) (20). Table 4 shows the comparison of the predicted and actual dye removal for 5 runs of the photocatalytic process.

**Table 4.** Metrics evaluated for the machine learning models.

|  |  |  |
| --- | --- | --- |
| **Run** | **Ypred** | **Yobs** |
| 1 | 49.95 | 49.18 |
| 2 | 50.37 | 49.51 |
| 3 | 40.49 | 36.72 |
| 4 | 27.36 | 24.92 |
| 5 | 38.61 | 36.72 |

Predicted values (RhB Removal) evaluated based on testing ANN model

The best condition for the heterogeneous photocatalysis process was set to [RhB] = 20 mg L-1, [nANA] = 2.75 g L-1, 25°C, yielding to 50.82% dye removal from aqueous solution. The kinetic curve of this run is shown in Figure 3.



**Figure 3**. Kinetic of dye photocatalytic degradation.

**Operational parameters:** [RhB] = 20 mg L-1 | [nANA] = 2.75 g L-1 | T = 25 ± 2 °C | pH 5.42 | monitoring time: 180 min, under visible light (202 W m-2).

According to Figure 3, the Langmuir-Hinshelwood was well fitted to experimental data, resulting in R2 and R2adj values very close to unity. Moreover, the low value evaluated for the error functions SAE, SSE, and ARE confirmed the accuracy of the linear model. Also, it can be observed about 50% of RhB removal (C C0-1 ~ 0.5) after 175 min of heterogeneous photocatalysis. The apparent rate constant equal to 0.0039 min-1 was reported, suggesting that about 0.39% of the dye molecules in solution are degraded after each minute (21). Thus, regarding this pattern, it is expected 70.2% RhB removal after 180 min. However, the actual observed value was 50.82%, which is expressively lower due to parallel reactions involved in the process or the generation of the intermediate products.

*Data correlation*

To investigate the correlation of each parameter with the response variable (Removal) and the rate constant (*k*), a Spearman correlation was performed, considering the non-gaussian distribution of the data. Thus, a heatmap generated from the Spearman method is presented in Figure 4.



**Figure 4.** Spearman correlation between input variables and output variables (*k*and removal).

According to Figure 4, it was observed that dye concentration ([RhB]) showed a moderate negative correlation to Removal and rate constant (*k*), suggesting that an increase in [RhB] should result in a decrease in the dye removal and *k*. Moreover, a moderate positive correlation between catalyst concentration ([nANA]) and the Removal and *k* was observed, which indicates that high values for these variables are expected with an increase in the [nANA]. The same behavior is encountered for pH, suggesting that RhB is degraded at a higher extent under neutral-slightly alkaline pH values (22). However, it is noticeable no correlation between [RhB] and [nANA], which agrees with experimental procedures. Figure 5 shows the feature importance study, in which the permutation score for each input parameter used in the ANN model is informed.



**Figure 5.** Feature importance study generated from ANN model.

According to Figure 5, all variables were significant to the ANN generalized model. It is evidenced by the permutation score, which was greater than 0.30. Greater the permutation score greater the influence of the feature (input variable) on the response. Thus, it is expected that slightly modifications in both [RhB] and [nANA] should generate different removal (%). Based on it, Figure 6 also suggests that abrupt/high alterations in the pH can affect significantly RhB removal, which was confirmed in experimental runs found in the literature (23).

## Conclusions

The present study proposed the use of an alternative low-silica nanozeolite as nanocatalyst for photocatalytic degradation of RhB from aqueous solution dye under visible light. Additionally, the development of a generalist model obtained from a machine learning study was proposed to predict dye removal. Thus, the generic predictive model was successfully developed using the ANN algorithm, whose structure 3:6:1 showed the best performance in the predictions of RhB removal by photocatalytic degradation. The ANN proposed was characterized by its high accuracy (R2 = 0.98-0.99 and RMSE < 5) and low probability of overfitting (R2> 0.80 for testing dataset). The ANN model predicted 50.37 ± 1.01 % of RhB degradation at [RhB] = 200 mg L-1, [nANA] = 2.75 g L-1 and pH 5.7-5.8 after 180 minutes under visible light. Moreover, data correlation and feature importance study confirmed that the dye/nanocatalyst concentration and pH significantly affect the dye removal and first-order rate constant (*k*). Therefore, this study confirms the potentiality of machine learning algorithms as useful tools for making predictions prior to experimental runs and as starting point for the scale-up of wastewater treatments with a focus on organic pollutants.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

## References

1. R. Kishor, D. Purchase, G.D. Saratale, R.G. Saratale, L.F.R Ferreira, M. Bilal, R. Chandra, R.N. Bharagava, J. Environ. Chem. Eng. **2021**, 9, 105012.
2. V. Gupta, V. Suhas. J. Environ. Manage. **2009**, 90, 2313-2342.
3. C.M. Magdalane, K. Kaviyarasu, J.J. Vijaya, B. Siddhardha, B. Jeyaraj, J. Kennedy, M. Maaza. J. Alloys Compd. **2017**, 727, 1324-1337.
4. S.B. Khan, M. Hou, S. Shuang, Z. Zhang Z, Appl Surf Sci **2017**, 400, 184-193.
5. R. Stanley, J.A. Jebasingh, M.S. Vidyavathy, P.K. Stanley, P. Ponmani, M.E. Shekinah, J. Vasanthi, Optik **2021**, 231, 166518.
6. T. Ennaert, J. Van Aelst, J. Dijkmans, R. De Clercq, W. Schutyser, M. Dusselier, D. Verboekend, B.F. Sels, Chem. Soc. Rev. 2016, 45, 584-611.
7. M.K. Shamsabadi, M. Behpour, Mater. Sci. Eng. B. **2021**, 269, 115170.
8. M. Yoldi, E. Fuentes-Ordoñez, S. Korili, A. GIL, Zeolite synthesis from industrial wastes. Microporous Mesoporous Mater **2019**, 287, 183-191.
9. Z.H. Jaffari, A. Abbas, S.-M. Lam, S. Park, K. Chon, E.-S. Kim, K.H. Cho, J. Hazard. Mater **2023**, 442, 120031.
10. X. Yin, M. Fallah-Shorshani, M., R. McConnell, S. Fruin, M. Franklin, Environ. Sci. Technol. **2020**, 54, 12860-12869.
11. A.R. Amani-Ghadim, M.S.S. Dorraji, Appl. Catal. B: Environ. **2015**, 163. 539-546.
12. Y. Zhang, Z. Leng, F. Zou, L. Wang, S.S. Chen, D.C.W. Tsang, J. Clean Prod. **2018**, 172, 686-695.
13. L.R. Oviedo, P.C.L. Muraro, G. Pavoski, D.C.R. Espinosa, Y.P.M. Ruiz, A. Galembeck, C.R.B. Rhoden, Environ. Sci. Pollut. Res. Int. **2022**, 29, 3794-3807.
14. S. Foteinis, A.G. Borthwick, Z. Frontistis, D. Mantzavinos, E. Chatzisymeon, J. Clean Prod. **2018**, 9, 1-12.
15. L.R. Oviedo, V.R. Oviedo, L.D.D. Nora, W.L. da Silva, Sep. Purif. Technol. **2023**, 315, 123712.
16. A. Altmann, L. Tolosi, O. Sander, T. Lengauer, Bioinformatics **2010**, 26, 1340-1347.
17. R.R. de Souza, M. Toebe, A.C. Mello, K.C. Bittencourt, Eur. J. Agron **2023**, 142, 126666.
18. T. Yin, X. Meng, S. Wang, X. Yao, N. Liu, L. Shi, Sep. Purif. Technol. **2022**, 280, 119634.
19. X. Zhu, D.C.W. Tsang, L. Wang, Z. Su, D. Hou, L. Li, J. Shang, J. Clean. Prod. **2020**, 273, 122915.
20. M.M. Najafabadi, T.M. Khoshgoftaar, F. Villanustre, J. Holt, J. Big Data **2017**, 4, 2-17.
21. R. Walsh, E. Martin, S. Darvesh, A method to describe enzyme-catalyzed reactions by combining steady state and time course enzyme kinetic parameters. Biochim Biophys Acta. **2010**, 1800, 1-5.
22. F. Soori, A. Nezamzadeh-Ejhieh, J. Mol. Liq. **2018**, 255, 250-256.
23. A. Nezamzadeh-Ejhieh, M. Karimi-Shamsabadi, Chem. Eng. J. **2013**, 228, 631-641.