MACEIÓ – ALAGOAS
EBA 15
15° ENCONTRO BRASILEIRO SOBRE ADSORÇÃO
15th Brazilian Meeting about Adsorption

Predictive Modeling of Enzymatic Immobilization of Lipases by Adsorption using Xgboost Algorithm aiming Esterification Reactions

Raimundo Vinicius Araujo Maia[a], Paula Maria Pereira Freire[a], Rebeca Teixeira de Queiroz Montenegro[a], João Lucas Souza de Oliveira[a], Luciana Rocha Barros Gonçalves[a*]

[a]*Federal University of Ceará, Department of Chemical Engineering, Fortaleza and Ceará*

Abstract

Enzymatic immobilization of lipases by adsorption can significantly enhance stability, preserve their properties, and enable the reuse of enzymes in industrial and laboratory processes, particularly in esterification. In industrial and laboratory applications, the XGBoost algorithm emerges as a promising tool for predictive modeling of the enzymatic immobilization of lipases by adsorption. This study is designed to predict the parameters associated with enzyme immobilization esterification, such as the substrate conversion rate, the number of reuse cycles, and the ideal temperature for the process. The methodology employed XGBoost, a boosting-based machine learning algorithm, to construct a predictive model that identifies complex relationships. Input variables included the name of the lipase, the immobilization support with its active chemical groups, the molarity of the buffer's ion, the immobilization time and temperature, and the substrate used to measure the enzymatic activity. The output variables were the number of reuse cycles, the optimal temperature for esterification, and the substrate conversion rate. The approach provided predictions about the ideal temperature to carry out esterifications with lipases immobilized by adsorption, the number of cycles supported with 60% of activity, and the substrate conversion efficiency. The predictions made by XGBoost underscore its effectiveness in handling complex data, thereby advancing our understanding and control of enzyme immobilization processes and contributing to the development of predictive methods in chemical and biotechnological engineering. These findings have practical applications in industrial and laboratory contexts, potentially improving the efficiency and cost-effectiveness of enzyme immobilization processes.

Keywords: Adsorption; Enzyme immobilization; XGBoost; Esterification; Predictive modeling.

1. Introduction

Enzymatic immobilization of lipases by adsorption is a consolidated technique that offers several advantages in using these enzymes in various industrial and laboratory processes. The fixation of enzymes on solid supports through adsorption makes it possible to improve the stability parameters of lipases, providing protection against variations in pH and temperature and making them more robust and efficient over time [1].

In addition to improving stability, immobilization by adsorption allows the reuse of lipases in multiple reaction cycles, thus making enzymatic processes economically viable [1].

Lipase immobilized by adsorption can be used in esterification solutions and is essential for chemical products such as esters, which are widely used in the food, cosmetic, and pharmaceutical industries. This process contributes to the sustainable and economically viable production of these compounds [1,2]. Immobilized enzymes, especially lipases, are indicated as effective tools in esterification processes, which involve the formation of esters from acids and alcohol. Due to the high specificity and catalytic efficiency of immobilized lipases, it is possible to achieve a high yield in the production of esters, operating under milder temperature and pressure conditions. In addition, using immobilized enzymes reduces the need for aggressive organic solvents, promoting a

more sustainable and economically advantageous process [3].

The need to improve and predict the behavior of immobilized enzymes has led to the adoption of predictive modeling techniques, such as the XGBoost algorithm [4].

This algorithm is particularly effective in handling large volumes of data and complex models. It offers excellent generalization capabilities and avoids the capture of noise and irrelevant features, known as overfitting. Due to its high computational efficiency, ability to analyze variable importance, and ability to handle missing data, XGBoost is widely used in regression and classification tasks [4].

In this context, the objective of this study is to predict the parameters related to enzyme immobilization esterification, such as the substrate conversion rate, the number of reuse cycles, and the optimal process temperature, using the XGBoost algorithm. The prediction will be made based on data extracted from articles available in the scientific literature.

2. Methods

The methodology used in this work was adapted from Chai et al (2021) [5] and Kang et al (2023) [6], with modifications. The dataset consists of 9 inputs: DOI of the paper, source of enzyme, support, group's support, ionic strength, immobilization time, and temperature and 3 outputs: number of cycles in which 60% of the activity of the enzyme is preserved, optimum temperature, and esterification conversion. The main sources of data collection were Scopus, PubMed, and Google Scholar.

After the data integration, the rows were preprocessed (filling of missing values and conversion of string into float values) and divided into train and test sets, the final size of the sets being 48 points for the non-augmented train set, 768 points for the augmented version, and 6 points for the test. In order to improve the generalization capacity of the XGBoost Regressor, gaussian noise injection was used as a data augmentation technique to expand the size of the training set. It is worth mentioning that the function utilized in this work can be described as:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)}{2\sigma^2}} \quad (1)$$

Where μ represents the mean of x, σ the standard deviation, and x the Gaussian noise to be generated and added to a virtual copy of the set during various iterations, followed by the addition to the original train set, which gives the final set used to feed the algorithm. To optimize the results of the predictions, 5-fold cross-validation and GridSearchCV [5] were employed to filter the hyperparameter values and give more accurate predictions. As the final step, the coefficient of determination (R2), Mean absolute error (MAE), and Root Mean Squared Error (RMSE) [5] were calculated for the train points and the absolute error for the test set.

3. Results & Discussion

First, the dataset size used to train the model was evaluated, and the esterification reaction conversion was selected as an output. The initial attempt to train the model without an augmented dataset resulted in relatively good adjusted R^2 but poor fit for the more distant values, from 60 to approximately 80% (see Figure 1). This unsatisfactory performance underscored the need for an augmented dataset, leading to the implementation of Gaussian noise injection to enhance the model performance, see Figure 2.

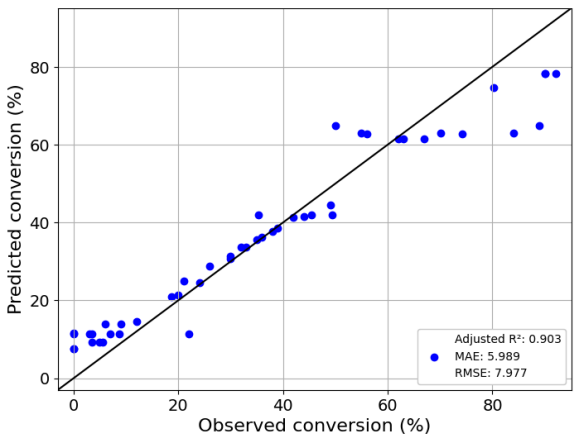

Fig. 1. Scatterplot for esterification reaction conversion (non-augmented train set).

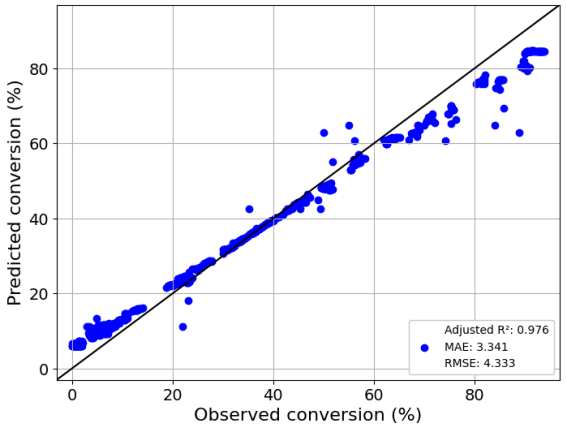

Fig. 2. Scatterplot for esterification reaction conversion (augmented train set).

The scatterplot and adjusted R^2 show that there is a strong correlation and explained variance of the output, see Figures 2 and 3. This is confirmed by the values of MAE and RMSE, which give relatively low errors for an average performance of predictions. The absolute error plot demonstrated that approximately 95% of the data contains an error of less than 10, the maximum deviation being 25 (% refers to the unit of .

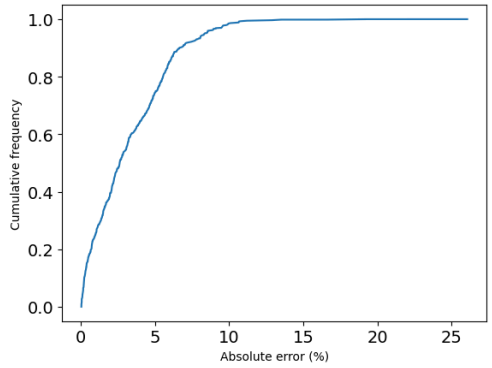

Fig. 3. Absolute error plot for esterification reaction conversion (%).

Since the data concerning the number of the days with 60% of enzymatic activity preserved was not present for many papers, it was not possible to obtain a reliable plot for this variable. This is due to the scarceness of information present in the literature, which shows the lack of interest for the measurement of this metric regarding the immobilization process for this biocatalyst.

As for the optimum temperature (Figures 4 and 5), the model could represent the experimental data since the predicted values are close to the observed values. The relation is robust even with some points of the central line with scarce quantities of data points.

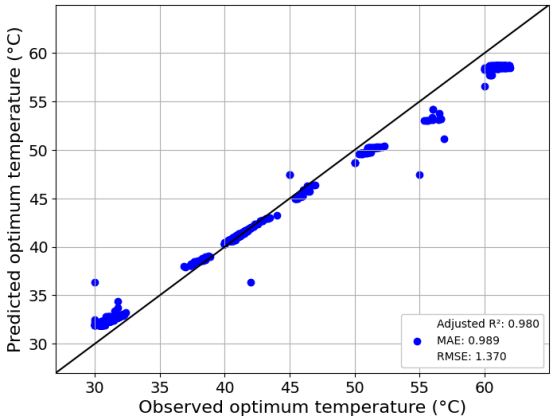

Fig. 4. Scatter Plot for optimum temperature (°C).

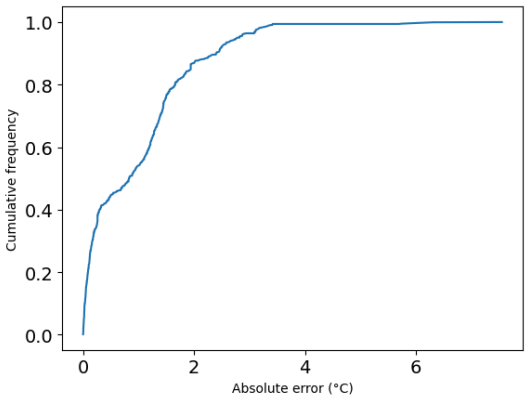

Fig. 5. Absolute error plot for optimum temperature (°C).

After the training process, the model's performance was evaluated using a separate set of data that was not included in the training. The results are documented in Table 1. Despite the differences between the predicted and observed results for some points, the model can indeed be utilized. However, its accuracy could be further enhanced by incorporating more real-world data from other works.

Table 1. Absolute errors for the test set.

Metric	Conversion (%)	Number of cycles (days)	Optimum temperature (°C)
Biggest absolute error (test)	29.24	58.26	1.49
Smallest absolute error (test)	4.05	0.52	0.26

4. Conclusion

The XGBoost algorithm was effectively employed to predict the effect of immobilization on esterification by lipases, using input variables such as the type of lipase, the immobilization support, and other experimental conditions. The developed predictive model successfully identified complex relationships. However, it's crucial to acknowledge that a more diverse expanded experimental set is required to further enhance the model's accuracy. These results provide valuable insights for improving industrial and laboratory processes and highlight the potential of XGBoost for driving advancements in enzyme immobilization.

Acknowledgments

The authors are grateful to CNPq (National Council for Scientific and Technological Development) and CAPES (Coordination of Superior Level Staff Improvement) for the financial support provided.

References

[1]FERNANDEZ-LAFUENTE, Roberto; ARMISÉN, Pilar; SABUQUILLO, Pilar; FERNÁNDEZ-LORENTE, Gloria; GUISÁN, José M.. Immobilization of lipases by selective adsorption on hydrophobic supports. Chemistry And Physics Of Lipids, [S.L.], v. 93, n. 1-2, p. 185-197, jun. 1998. Elsevier

[2]JURADO-DAVILA, Vanessa; OSHIRO, Gabriel Pollo; ESTUMANO, Diego Cardoso; FÉRIS, Liliana Amaral. Immobilization of Marbofloxacin for Water Treatment by Adsorption in Batch Scale and Fixed-Bed Column: applying of monte carlo bayesian modeling. Industrial & Engineering Chemistry Research, [S.L.], v. 63, n. 22, p. 9976-9987, 23 maio 2024.

[3]Basso A., Serban S. Industrial Applications of Immobilized Enzymes—A Review. Molecular Catalysis, 479, 110607, 2019

[4]Wang Q, Zou X, Chen Y, Zhu Z, Yan C, Shan P, Wang S, Fu Y. XGBoost algorithm assisted multi-component quantitative analysis with Raman spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy. V 323, p.124917, 2024.

[5]CHAI, Milton; MORADI, Sina; ERFANI, Eila; ASADNIA, Mohsen; CHEN, Vicki; RAZMJOU, Amir. Application of Machine Learning Algorithms to Estimate Enzyme Loading, Immobilization Yield, Activity Retention, and Reusability of Enzyme–Metal–Organic Framework Biocatalysts. Chemistry Of Materials. v. 33, n. 22, p. 8666-8676,2021.

[6]KANG, Zhongming; FENG, Lianfang; WANG, Jiajun. Optimization of a Gas–Liquid Dual-Impeller Stirred Tank Based on Deep Learning with a Small Data Set from CFD Simulation. Industrial & Engineering Chemistry Research. V. 63, p. 843-855, 2023.