

A Comparison of Machine Learning Approaches in Predicting Viscosity for Partially Hydrolyzed Polyacrylamide Derivatives[☆]

Comparação de Abordagens de Aprendizado de Máquina na Predição da Viscosidade para derivados de Poliacrilamida Parcialmente Hidrolisada

Kelly Cristine da Silveira^{1,†}, Matheus Henrique Silva Siqueira¹, João Matheus Ramos Gama¹, Jonathan Nogueira Gois², Cláudio Fabiano Motta Toledo³, Antônio J. Silva Neto¹

¹ State University of Rio de Janeiro, Polytechnic Institute, Nova Friburgo, RJ, Brazil

² Federal Center for Technological Education Celso Suckow da Fonseca, Nova Friburgo, RJ, Brazil

³ University of São Paulo, Institute of Mathematics and Computer Sciences, São Carlos, SP, Brazil

[†]Corresponding author: kelly.dasilveira@iprj.uerj.br

Abstract

Partially hydrolyzed polyacrylamides (HPAM) are widely used to modulate the viscosity of formulations. The appropriate application of a viscosity model can facilitate the idealization of new macromolecules and contribute to a better understanding of the structure-property relationship. In the present study, machine learning approaches, Multiple Linear Regression (MLR) and Random Forest (RF), were compared to model the viscosifying effect of HPAM derivatives, based on their chemical composition and concentration in an aqueous solution. The evaluated data come from a previous experimental study, which explores a post-synthetic polymer modification methodology. The relative importance of the variables was investigated, determining the features with the greatest influence on viscosity, including variations in chemical composition, with emphasis on the more hydrophobic groups (C7 and C12). The accuracy of the models was evaluated using statistical criteria, the coefficient of determination (R^2) and the Root Mean Square Error (RMSE). The Random Forest approach outperformed Multiple Linear Regression, with values of 0.97 and 0.30 for R^2 and RMSE, respectively, compared to 0.83 and 0.67 for Multiple Linear Regression. Applying the Random Forest model, it was possible to generate a set of hypothetical macromolecules, with potential viscosifying effects. These macromolecules were idealized focusing on mixed compositions of C7 and C12 with a maximum structural variation of 10 mol%. Additionally, this structural mapping provided insights for designing promising polymers by inserting cyclic structures, such as CYCLOPROP, which could overcome the solubility limitation observed in the literature.

Keywords

Partially Hydrolyzed Polyacrylamide • Machine Learning • Viscosity

[☆] This article is an extended version of the work presented at the Joint XXV ENMC National Meeting on Computational Modeling, XIII ECTM Meeting on Science and Technology of Materials, 9th MCSul South Conference on Computational Modeling and IX SEMENGO Seminar and Workshop on Ocean Engineering, held in webinar mode, from October 19th to 21th, 2022

Resumo

Poliacrilamidas parcialmente hidrolisadas (HPAM) são amplamente utilizadas para modular a viscosidade de formulações. A aplicação adequada de um modelo de viscosidade pode facilitar a idealização de novas macromoléculas e contribuir para um melhor entendimento da relação estrutura-propriedade. No presente estudo, abordagens de aprendizado de máquina, Regressão Linear Múltipla (MLR) e Floresta Aleatória (RF), foram comparadas para modelar o efeito viscosificante de derivados HPAM, com base em sua composição química e concentração em uma solução aquosa. Os dados avaliados provêm de um estudo experimental anterior, que explora uma metodologia de modificação pós-sintética de polímeros. A importância relativa das variáveis foi investigada, determinando as características com maior influência na viscosidade, incluindo variações na composição química, com destaque para os grupos mais hidrofóbicos (C7 e C12). A acurácia dos modelos foi avaliada por meio de critérios estatísticos, coeficiente de determinação (R^2) e Root Mean Square Error (RMSE). A abordagem Floresta Aleatória superou a Regressão Linear Múltipla, com valores de 0,97 e 0,30 para R^2 e RMSE, respectivamente, em comparação com 0,83 e 0,67 para Regressão Linear Múltipla. Aplicando o modelo Floresta Aleatória, foi possível gerar um conjunto de macromoléculas hipotéticas, com potenciais efeitos viscosificantes. Esses polímeros foram idealizados com foco em composições mistas de C7 e C12, com variação estrutural máxima de 10 mol%. Adicionalmente, o mapeamento estrutural forneceu subsídios para o *design* de polímeros promissores através da inserção de estruturas cíclicas, como CYCLOPROP, o que pode superar a limitação de solubilidade observada na literatura.

Palavras-chave

Poliacrilamida parcialmente hidrolisada • Aprendizado de máquina • Viscosidade

1 Introduction

Water-soluble polymers are heavily used as viscosity modifiers. Polyacrylamides, especially partially hydrolyzed polyacrylamide (HPAM), are promising viscosifiers in various applications in the oil industry, such as hydraulic fracturing, drilling fluids, and formulations used for advanced oil recovery [1]. Still, they receive attention in several industrial sectors, being applied in formulations of paints, food, water purification, pharmaceuticals, and cosmetics. However, polyacrylamide-based polymers are sensitive to mechanical degradation in environments with high temperature, shear, or in the presence of electrolytes [2], making a great effort necessary for the synthesis of their derivatives, aiming to overcome this limitation, with emphasis on the known hydrophobically modified water-soluble polymers (HMWSP).

HMWSP are amphiphilic copolymers consisting of a large hydrophilic portion in the chain and a small hydrophobic portion, with an incorporation limit of less than 5 mol% [3]. The hydrophobic groups incorporated into the macromolecule can promote a significant increase in viscosity, as they tend to associate, pushing away water molecules and generating intra or intermolecular aggregates in solution [2]. In this context, the ability to control the addition of functional groups is essential to maximize the effect on viscosity and minimize the loss of performance observed in polyacrylamides, specifically, the partially hydrolyzed polyacrylamides.

Faced with this challenge, an alternative synthetic route was evaluated by Da Silveira *et al.* [4]. It is a synthesis method based on the coupling reaction with carbodiimide, used for the post-synthetic modification of polyacrylamides in water. In this protocol, the carboxylic acid groups present in the commercial HPAM polymer are activated for subsequent chemical modification, generating a wide variety of copolymers and terpolymers with minimal variation in molar mass and dispersity (\bar{M}_w/\bar{M}_n) between samples. The modification occurs along the chain, avoiding fluctuation in the final polymer composition, being an essential aspect in the study of the correlation between structure and property of interest.

Although a large number of new structures have been previously synthesized [4], the correlation between the incorporated clusters must be further investigated. In addition, the application of these polymers is limited due to their viscosifying performance, which makes a model of great relevance for mapping the possible structures to be explored, according to the viscosity range required for their application. It is also important to point out that the literature presents studies with few structural variations or presents biased correlations, in which the macromolecules have very different molar masses, limiting the comparative approach of the material properties. Thus, the dataset evaluated in this work comes from an experimental method that aims to minimize this problem, offering the advantage of fine-tuning the structure of the polymer.

From the computational point of view, the literature shows that predictive modeling may be able to relate experimental data quickly and completely, making it an attractive strategy to accelerate the discovery of new

materials or even facilitate the screening of new structures [5, 6]. Supervised machine learning methods can perform prediction and inference tasks without limiting theoretical assumptions to the experiment. The main advantage consists in the possibility of performing the inference of new data quickly. It is worth mentioning that the concern with data generation is fundamental, where representative data guarantee the consistency of the model [7]. However, compared to experimental curation techniques, machine learning proves to be a more economical and powerful tool when mapping datasets, which may not be systematically varied, into an algorithm with a great capacity to improve prediction efficiency [5].

With wide use, Multiple Linear Regression (MLR) algorithms are used to model how certain characteristics of the phenomenon infer a certain property or result. Once the relationships between the dependent variables are established, the technique can be used as a tool for prediction. For example, a multiple linear regression model was applied by Najafabadi and Soares [8] to evaluate the effect of polymer architecture on the flocculation and dewatering of tailings from tar sands with the polymer EPDM-g-PMA. The performance of the copolymers was measured by the initial sedimentation rate, capillary suction time, supernatant turbidity, and sediment solids content. However, the MLR approach only assumes linear relationships between predictors and dependent variables, which can lead to prediction errors. On the other hand, Decision Tree-based models are gaining popularity in the literature, among them Random Forest (RF), which tolerates multicollinearity, that is, different data with the same predicted characteristic value. This is due to the ability of this algorithm to subsample the training set into small sets and then apply the decision tree technique. In this way, it overcomes some limitations imposed by the MLR approach [9, 10].

The Random Forest (RF) algorithm is a non-parametric regression method that uses a combination of different models, known as *bagging*, to make predictions. Developed by Breiman [9] for classification and regression, it is robust and capable of evaluating variable importance and selecting important features. Despite dealing with many variables and limited observations, the algorithm delivers high accuracy and is not affected by multicollinearity. In environmental modeling [10, 11] and the development of polymeric materials [12, 13], the RF approach has been widely applied and shown to perform well in classification and prediction tasks.

In this context, the current study aims to apply a prediction model using machine learning to estimate the viscosity in an aqueous solution of partially hydrolyzed polyacrylamide derivatives, reducing the laboratory demand for structural mapping of the polymeric family. A comparison of two approaches, Multiple Linear Regression (MLR) and Random Forest (RF), is presented. This preliminary work not only showcases the use of machine learning algorithms in designing new polymers but also offers a fast and cost-effective mapping of the structure-viscosity relationship for the polyacrylamide derivatives studied.

2 Materials and Methods

2.1 Experimental data

The polymers evaluated in this study were obtained through post-synthetic modification, a versatile and non-toxic technique [4]. The precursor polymer (HPAM) was modified by adding 8 segments of the following amines in 2 and 20 mol%: isopropylamine (isoprop), cyclopropylamine (cycloprop), cyclobutylamine (cyclobut), cyclopentylamine (cyclopent), cycloheptylamine (cyclohept), cyclooctylamine (cyclooct), heptylamine (C7), and dodecylamine (C12). Viscosity measurements were performed using a *Tuning Fork Vibro Viscometer, A&D Company (SV-1A)* at a natural vibration frequency of 30 Hz. The polymer solutions were made by dissolving the dried polymers in distilled water at room temperature and evaluating concentrations from 0.3 to 10 g/L.

2.2 Data Organization

The experimental data was organized based on the final polymeric structure, which was determined by the repetition of each monomeric group, represented by the degree of polymerization (DP). The DP value was used to indicate the average number of times each group appeared in the macromolecule, referred to as a *block*. For clarity, Fig. 1 shows the diagram of the precursor polymer modified with 2 mol% of isopropylamine, poly(acrylamide-acrylic acid-*N*-isopropylacrylamide), denoted as PAM-AA-ISOPROP-2. The simplified composition of this polymer, represented by its blocks, is [1688, 375, 41]. Since the amide block (acrylamide) is constant in all evaluated structures, it was removed from the dataset matrix, without affecting the observed correlation. It is important to note that the term "block" used in this work does not refer to block copolymers. All polymers evaluated in this study are statistical copolymers.

The dataset consists of 208 observations for 16 different polymeric structures, with their viscosity in water evaluated at 13 concentrations ranging from 0.3 to 10 g/L. The composition of these structures is based on the incorporation of 8 different chemical structures (segments) derived from primary amines, added individually to the activated polymer, and presented as a *modified block* in Table 1. The modification percentage was 2 and 20 mol%, translated to 41 and 416 blocks, respectively. The block related to the acrylic acid monomer is referred to as the *AA block* and it encompasses the functional group of carboxylic acid that has been chemically activated. For each of the 208 observations, 10 input parameters were used to predict the viscosity property, including the *AA block*, the *modified block* (with 8 possible structures), and concentration.

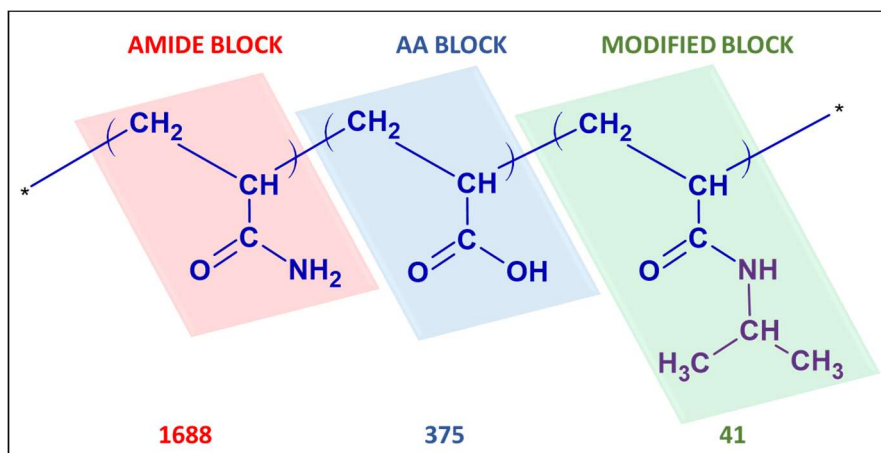


Figure 1: Schematic representation of the PAM-AA-ISOPROP-2 polymer structure as a function of the average number of *blocks* in the macromolecule.

Table 1: Summary of parameters.

Parameters	Value range			
Composition	<i>AA block</i>	0	375	416
	<i>Modified block</i>	416	41	0
Concentration [g/L]	0.3		10	
Viscosity [mPa.s]	1.0		13.2	

2.3 Development and comparison of prediction models

All data analysis and the application of machine learning algorithms, MLR and RF, were performed using *Python* with the following libraries: *Pandas*, *Seaborn*, *NumPy*, *Scikit-learn*, and *Matplotlib*. A preliminary exploratory data analysis was performed to examine the distribution of the data. The correlation between variables was calculated to understand the strength of the relationship, with a coefficient in the interval -1 to 1.

For the application of machine learning algorithms, no pre-processing was done on the experimental data. The same dataset was used for both MLR and RF algorithms, containing 208 data points with 80% used for training and 20% for testing, randomly allocated. The parameters are summarized in Table 1. The performance of the models was evaluated based on the metrics of coefficient of determination (R^2) and Root Mean Squared Error (RMSE). RMSE, defined by Eq. (1), is a metric used to evaluate the accuracy of the model by calculating the root mean square of errors between the observed values (actual) and the predicted values (hypothesis).

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (1)$$

The R^2 coefficient is a dimensionless statistical value that varies from 0 to 1. It measures the degree of fit of a generalized linear statistical model, with values closer to 1 indicating a better fit. It is calculated using Eq. (2), where \bar{y} is the average of the experimental values for n observations.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\bar{y} - y_i)^2} \quad (2)$$

To determine the predictive capacity, the following configuration was adopted for the training and validation of the models:

- (i) Random division of the dataset into training (80%) and test (20%);
- (ii) Adjustment of model parameters using the training dataset;
- (iii) Viscosity prediction applying the test data to the trained model;
- (iv) Calculation of metrics for model evaluation (R^2 and RMSE) using predicted and observed data from the test dataset.

2.4 Mapping of new polymeric structures

The final stage of this study consists of applying the implemented model, with higher accuracy, to predict the viscosity of hypothetical macromolecules. Furthermore, to experimentally validate the model, 6 experimental data were examined, and their parameters are presented in Table 2. The hypothetical polymeric structures, amounting to a total of 30, were designed based on the importance levels of the predictors. The parameters of these structures are consolidated in Table 3.

Table 2: Summary of parameters of the polymeric structures evaluated experimentally.

Concentration [g/L]	0.1	0.2	0.4	8
	PAM-AA-C7-2 ^a			
Polymers	PAM-AA-C12-2 ^a	HPAM ^c	HPAM ^c	HPAM ^c
	PAM-C7-20 ^b			

^a Polymers with 2 mol% modification; simplified standard configuration [1688, 375, 41]

^b Polymer with 20 mol% modification; simplified standard configuration [1688, 0, 416]

^c Partially hydrolyzed polyacrylamide with simplified standard configuration [1688, 416, 0]

Table 3: Summary of parameters of hypothetical polymeric structures.

Parameters	Value range		
Chemical modification (mol%)	<i>simple</i> composition (C12)	1	14
	<i>mixed</i> composition (C7 + C12)	1+1	10+10
	<i>mixed</i> composition (cycloprop + isoprop + C12)	1 + 1 + 1	1 + 1 + 5
Concentration [g/L]	10		

3 Results and discussion

The machine learning algorithms, Multiple Linear Regression and Random Forest were compared in this work and applied to a series of polymers synthesized through post-synthetic modification of the commercial product partially hydrolyzed polyacrylamide (HPAM). All the polymers had very similar molar masses but differed in composition due to the incorporation of a variety of chemical structures into the polymeric chain. The commercial sample HPAM was used as a precursor polymer and consists of a synthetic copolymer made of acrylamide monomer (~80 mol%) and acrylic acid in partial sodium salt form. By replacing 2 mol% of the carboxylate groups with various primary amines, a significant change in viscosity was observed. In aqueous media, hydrophobic groups tend to associate to reduce their contact with water, resulting in the formation of hydrophobic domains. With increasing concentration,

these segments can associate, leading to the formation of temporary intermolecular networks, and thus an increase in viscosity, depending on the chemical composition of the macromolecule [4].

The viscosity profile of the evaluated polymers is shown in Fig. 2(a), where the experimental viscosity data are displayed as a function of polymeric concentration. A unique profile of viscosity increase can be observed for each polymer composition, but the viscosity ranges at the same evaluated concentration exhibit a pronounced response to the chemical structure of the modified macromolecule and its modification percentage (mol%). To better observe the viscosifying effect of the chemical modification, Fig. 2(b) presents the separation of data based on the chemical modification percentage, indicated indirectly by the number of blocks related to the acrylic acid monomer, AA block. *AA block 0* represents polymeric structures with 20 mol% modification, which is the threshold modification for the chosen experimental protocol; *AA block 375* corresponds to polymers with 2 mol% modification; and *AA block 416* represents the precursor polymer, HPAM.

Fig. 2 (b) highlights the dataset referring to polymers with 2 mol% chemical modification (in red) as the most important in the rheological response. Furthermore, the polymers observed with a decrease in viscosity values (in blue) are polymers with the maximum modification percentage (20 mol%). The modified block with the greatest effect on the viscosity of the polymeric solution was C12 (PAM-AA-C12-2), as expected because it is a group with an intense hydrophobic character. The effect of increasing polymeric concentration on the hydrophobic association, and also on viscosity, is greater due to increased intermolecular interactions. Two distinct linear profiles can be observed for each modifier at 2 mol%, especially C12, and C7, which may be related to the transition in the concentration regime, called critical concentration (C^*). This transition occurs at lower concentrations for the C12-polymer dataset, as intermolecular interactions are facilitated. The change in the interaction profile of the macromolecule in water is known and is related to the interaction behavior of the polymeric coils, which in a semi-diluted regime exhibit stronger interactions, leading to an increase in observed viscosity.

The correlation matrix of the evaluated features, shown in Fig. 3, highlights the main predictors related to the increase in viscosity in the solution. The importance of the concentration is observed, with the semi-diluted regime presenting the highest values of viscosity regardless of the macromolecule observed. However, as already discussed, the C12 block is highlighted in the viscosifying effect, as well as the AA block, referring to the acrylic acid monomer. The other modifiers have a reduced ability to increase viscosity, due to the low degree of hydrophobicity, however, they are structures that cannot be discarded, as they have potential applications particularly in the inhibition of gas hydrates by polymers [14, 15, 16].

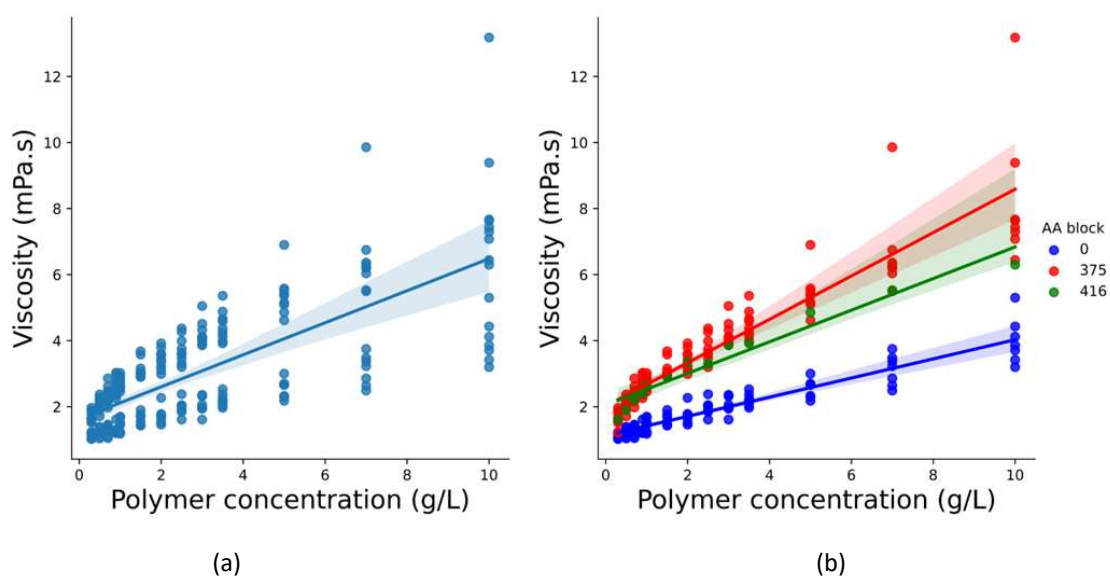


Figure 2: (a) Viscosity profile of the set of macromolecules studied. (b) Viscosity profile for 2 mol% modified polymers (red, 375); 20 mol% modified polymers (blue, 0) and polymer precursor, HPAM (green, 416).

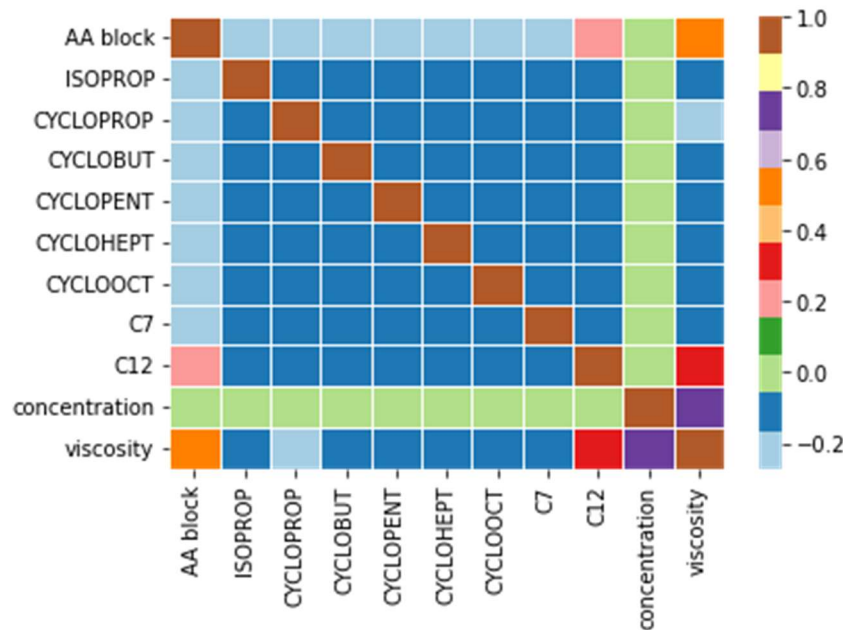


Figure 3: Correlation matrix of the studied features.

The challenge of solubility in water for post-modified macromolecules is particularly evident when incorporating highly hydrophobic groups. In a previous study by Da Silveira *et al.* [4], the modification of HPAM polymer with C12 at a concentration of 20 mol% led to insolubility, preventing the determination of viscosity for this sample. This experimental limitation highlights the critical importance of carefully selecting the incorporation of hydrophobic groups and optimizing the structure to achieve a balance between viscosity and solubility. Although the literature suggests a maximum level of hydrophobic modification of 5 mol% [3], this study challenged this limitation by using machine learning to map the structure-property relationship. Through the evaluation of different combinations of modified blocks, the machine learning models aimed to design new macromolecules with optimized viscosity while retaining water solubility, a crucial factor for their successful application.

Using the data set reserved for testing (20%), the models implemented were evaluated. For the MLR approach, the 10 predictors entered Eq. (3), with an R^2 of 0.83 and RMSE of 0.67 mPa.s, where Y is the variable of interest (viscosity), C is the intercept with a value of 2.3 and the values of β and X are shown in Table 4.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_{10} X_{10} + C \quad (3)$$

Table 4: Equation values obtained by MLR algorithms.

Variable	Feature	Coefficient	Value
X_1	AA	β_1	3.63×10^{-4}
X_2	ISOPROP	β_2	-3.45×10^{-3}
X_3	CYCLOPROP	β_3	-5.27×10^{-3}
X_4	CYCLOBUT	β_4	-4.25×10^{-3}
X_5	CYCLOPENT	β_5	-5.00×10^{-3}
X_6	CYCLOHEPT	β_6	-4.13×10^{-3}
X_7	CYCLOOCT	β_7	-3.80×10^{-3}
X_8	C7	β_8	-4.01×10^{-3}
X_9	C12	β_9	2.95×10^{-2}
X_{10}	Concentration	β_{10}	5.06×10^{-1}

For the RF algorithms, the model was also built using the entire set of predictors and had an overall performance of $R^2 = 0.97$ and RMSE = 0.30 mPa.s. In direct comparison with the MLR model, RF provides a significant performance boost when using the same predictors. Figure 4 presents a visual comparison of the MLR and RF

models for the 42 predictions made. An inspection of the values reserved for testing shows that their occurrences were randomly distributed throughout the dataset, running through the diluted and semi-diluted regimes.

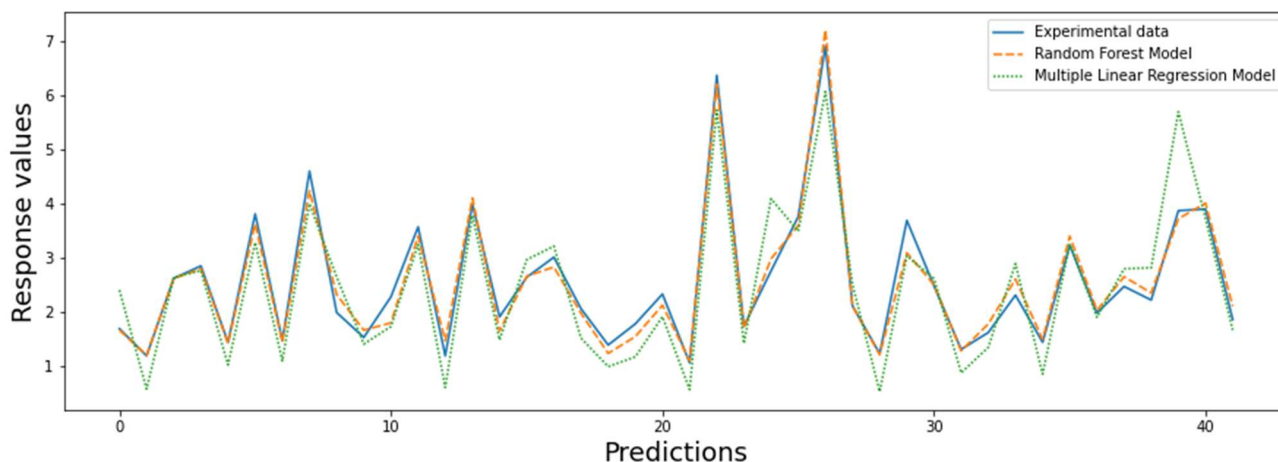


Figure 4: Comparison of models in predicting response values. Multiple Linear Regression Model - $R^2 = 0.83$ and RMSE = 0.67 mPa.s; Random Forest Model - $R^2 = 0.97$ and RMSE = 0.30 mPa.s.

The subsequent stage in the analysis involved the evaluation of a set of hypothetical polymers, leveraging the feature importance determined from Fig. 5, to gain a deeper understanding of the influence of structural characteristics on viscosity variation using RF algorithms. This underscores the significance of creating a mapped viscosifying profile, particularly in industrial sectors where a reduction in the rheological impact of the final product is deemed desirable.

The experimental data presented in this study is limited in terms of the number of polymeric structures examined, which may restrict the ability to accurately predict the viscosity values of novel hypothetical polymers. As such, experimental validation was conducted to determine the best model for predicting viscosity, with the RF model once again demonstrating superior performance. Conversely, the MLR model was found to inadequately generalize the rheological behavior of macromolecules in new datasets and hypothetical polymer compositions. While the MLR model exhibits a high R^2 value for predicting response variables, its predictive capabilities are hindered by limitations in accounting for chemical complexity and the influence of individual characteristics on viscosity variation. The inadequacies of the MLR model are evident in the evaluation of hypothetical polymers, where it fails to account for the well-known behavior of highly hydrophobic groups like C7 and C12, which can lead to a significant reduction in viscosity values as the maximum modification is approached [4]. In contrast, the RF model is well-equipped to account for such characteristics and can establish a modification threshold to optimize viscosity in hypothetical polymers.

In previous analyses (Fig. 3), the significance of C12 and AA blocks has already been established. In Fig. 5, the presence of modified block C7, which exerts a weaker influence due to its hydrophobic character, is observed, along with the introduction of blocks ISOPROP and CYCLOPROP in a relatively minor capacity. Despite their limited hydrophobic contribution, these blocks have the potential to optimize the viscosifying properties of a macromolecule without impeding solubility in water. This may be achieved through their synergistic interaction with the primary predictors already identified, notwithstanding their relatively weak hydrophobicity. In this context, it is important to note that experimental validation of the model at this stage requires new data with different structural profiles, as the training data do not account for the combination of modified blocks in mixed macromolecules. However, hypothetical polymers were evaluated, and the model suggested a possible polymeric modification of up to 10 mol% using C7 and C12 blocks (5 mol% of each). Beyond an 11 mol% modification with C12, the model indicates a marked reduction in the response variable, as expected, which may be attributed to the prevalence of intramolecular interactions, resulting in contraction of the polymeric chain and consequently, a decline in viscosity.

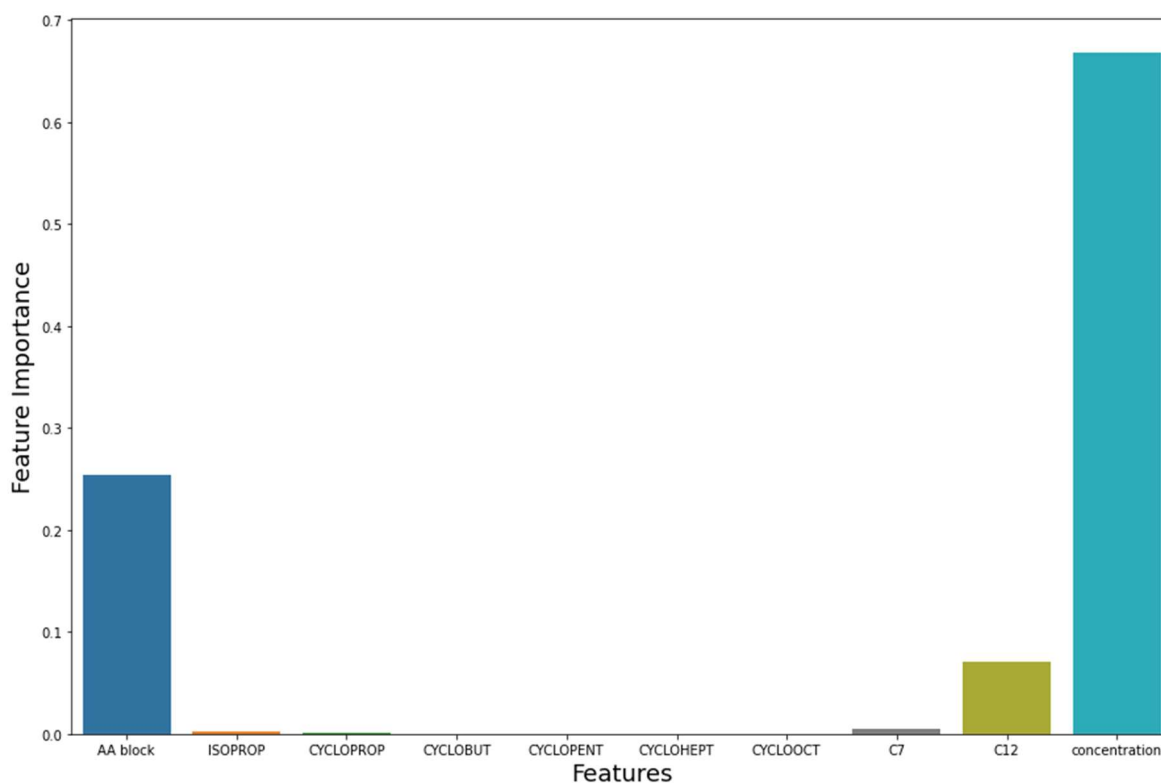


Figure 5: Level of importance of the features studied using RF algorithms.

Moreover, the RF model exhibits the potential for mixed macromolecules, where the combination of modified blocks C12, ISOPROP, and CYCLOPROP results in maximum viscosity values. These hypothetical structures represent novel polymers with a general structure of PAM-AA-ISOPROP-CYCLOPROP-C12, demonstrating the incorporation of structures that may enhance the solubility parameters of Hansen [17]. Specifically, cyclic chains contribute to both dispersion forces (δ_D) and hydrogen bonds (δ_H), indicating a more hydrophilic character for the block. As a result, the structural mapping provided by the Random Forest model introduces innovative structures with potential viscosifying effects, overcoming the solubility limitations observed in the literature for hydrophobically modified macromolecules.

4 CONCLUSIONS

The present study reports the use of machine learning approaches in the implementation of viscosity prediction models for partially hydrolyzed polyacrylamide derivatives. Featuring the best metrics, the model using Random Forest algorithms offers structures with a potential viscosifying profile, highlighting the incorporation of blocks C7, and C12, but also the cyclic structure, like CYCLOPROP, with reduced hydrophobic character. The findings of this study offer valuable insight into the structure-property relationship of HPAMs and their viscosifying effect, providing a foundation for the optimization of their synthesis and applications across diverse industrial sectors.

Acknowledgments

The authors are grateful for the support received by the Coordination for the Improvement of Higher Education Personnel Brazil (CAPES) - funding code 001; Research Support Foundation of the State of Rio de Janeiro (FAPERJ) and the National Council for Scientific and Technological Development (CNPq).

References

- [1] E. F. Lucas, L. S. Spinelli, and C. N. Khalil, “Polymers Applications in Petroleum Production”. In *Encyclopedia of Polymer Science and Technology*, 5th ed., 2015, pp. 01-49. Available at: <https://doi.org/10.1002/0471440264.pst641>.
- [2] Y. Feng, L. Billon, B. Grassl, A. Khoukh, and J. François, “Hydrophobically associating polyacrylamides and their partially hydrolyzed derivatives prepared by post-modification. 1. Synthesis and characterization,” *Polymer*, vol. 43, pp. 2055-2064, 2002. Available at: [https://doi.org/10.1016/S0032-3861\(01\)00774-1](https://doi.org/10.1016/S0032-3861(01)00774-1).
- [3] W. Kuang, and C. W. Gao, “Synthesis and Characterization of Novel Twin-Tailed Hydrophobically Associated Copolymers and Their Applications to Cr(III) Removal from Aqueous Solutions,” *Journal of Applied Polymer Science*, vol. 131, 41028, 2014. Available at: <https://doi.org/10.1002/app.41028>.
- [4] K. C. Da Silveira, Q. Sheng, W. Tian, E.F. Lucas, and C. D. Wood, “Libraries of modified polyacrylamides using post-synthetic modification,” *Journal of Applied Polymer Science*, vol. 132, no. 47, 2015. Available at: <https://doi.org/10.1002/app.42797>.
- [5] J. N. Kumar, Q. Li, K.Y. Tang, T. Buonassisi, A. L. Gonzalez-Oyarce, and J. Ye, “Machine learning enables polymer cloud-point engineering via inverse design,” *Npj Computational Materials*, vol 5, pp. 1-6, 2019. Available at: <https://doi.org/10.1038/s41524-019-0209-9>.
- [6] L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth, and R. Ramprasad, “Polymer informatics: Current status and critical next steps,” *Materials Science and Engineering: R: Reports*, vol. 144, 100595, 2021. Available at: <https://doi.org/10.1016/j.msere.2020.100595>.
- [7] S. Wu, Y. Kondo, M. A. Kakimoto, M. A. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, and R. Yoshid, “Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm,” *Npj Computational Materials*, vol 5, pp. 1-6, 2019. Available at: <https://doi.org/10.1038/s41524-019-0203-2>.
- [8] Z. R. Najafabadi, and J. B. Soares, “Flocculation and dewatering of oil sands tailings with a novel functionalized polyolefin flocculant,” *Separation and Purification Technology*, vol. 274, 119018, 2021. Available at: <https://doi.org/10.1016/j.seppur.2021.119018>.
- [9] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001. Available at: <https://doi.org/10.1023/A:1010933404324>.
- [10] G. Y. Oukawa, P. Krecl, and A. C. Targino, “Fine-scale modeling of the urban heat island: a comparison of multiple linear regression and random forest approaches,” *Sci. Total Environment*, vol. 815, 152836, 2022. Available at: <https://doi.org/10.1016/j.scitotenv.2021.152836>.
- [11] I. Ouedraogo, P. Defourny, and M. Vanclooster, “Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale,” *Hydrogeology Journal*, vol. 27, pp. 1081–1098, 2019. Available at: <https://doi.org/10.1007/s10040-018-1900-5>.
- [12] Y. Ueki, N. Seko, and Y. Maekawa, “Machine learning approach for prediction of the grafting yield in radiation-induced graft polymerization,” *Applied Materials Today*, vol. 25, 101158, 2021. Available at: <https://doi.org/10.1016/j.apmt.2021.101158>.
- [13] M. Kishino, K. Matsumoto, Y. Kobayashi, R. Taguchi, N. Akamatsu, and A. Shishido, “Fatigue Life Prediction of Bending Polymer Films using Random Forest,” *International Journal of Fatigue*, 107230, 2022. Available at: <https://doi.org/10.1016/j.ijfatigue.2022.107230>.

- [14] K. C. Da Silveira, Q. Sheng, W. Tian, C. Fong, N. Maeda, E. F. Lucas, and C. D. Wood, “High throughput synthesis and characterization of PNIPAM-based kinetic hydrate inhibitors,” *Energy & Fuels*, vol. 188, pp. 522–529, 2017. Available at: <https://doi.org/10.1016/j.fuel.2016.10.075>.
- [15] J. Park, K. C. Da Silveira, Q. Sheng, C. D. Wood, and Y. Seo, “Performance of PNIPAM-based Kinetic Hydrate Inhibitors for Nucleation and Growth of Natural Gas Hydrates,” *Energy & Fuels*, vol. 31, no. 3, pp. 2697–2704, 2017. Available at: <https://doi.org/10.1021/acs.energyfuels.6b03369>.
- [16] I. R. Ferreira Filho, K. C. Da Silveira, and A. J. Silva Neto, “Tecnologia de Hidrato de Gás: Modelagem Computacional para a Etapa de Crescimento do Hidrato,” *VETOR - Revista De Ciências Exatas E Engenharias*, vol. 31, no. 1, pp. 23–42, 2021. Available at: <https://doi.org/10.14295/vetor.v31i1.13164>.
- [17] C. M. Hansen, (2007), “Hansen Solubility Parameters: A User's Handbook”, 2nd ed. Boca Raton, United States: CRC Press, 2007. Available at: <https://doi.org/10.1201/9781420006834>.