Prediction of Hydrophobic Properties in Biopolymer-Based Coatings via Formulation Data Modelling

Anja Verbič,[1] Vuk Martinović,[1] Rubén Saborido,[2] Antonio Borrego,[3] Mariano Luque,[3] Uroš Novak,[1] Blaž Likozar[1], Blaž Stres[1]

[1] National Institute of Chemistry, Department of Catalysis and Chemical Reaction Engineering, Ljubljana Slovenia
anja.verbic@ki.si, vuk.martinovic@ki.si, uros.novak@ki.si, blaz.likozar@ki.si, blaz.stres@ki.si
[2] University of Málaga, Institute for Software Engineering and Software Technology "Jose Maria Troya Linero", Málaga, Spain
rsain@uma.es
[3] University of Málaga, Department of Applied Economics (Mathematics), Málaga, Spain
antoniobo@uma.es, mluque@uma.es

The urgent need to replace environmentally persistent and toxic per- and polyfluoroalkyl substances (PFAS) in hydrophobic coatings, has driven the exploration of renewable biopolymers as sustainable alternatives. Biopolymer-based coatings, derived from chitosan, cellulose or starch, are promising alternatives, but optimising their formulations to achieve targeted performance remains challenging. The traditional experimental approaches are time-consuming and resource-intensive. This study integrated a simple linear regression (SLR) technique based on ordinary least squares (OLS) to model the relationship between the formulations` composition and resultant surface water contact angles (WCA) achieved when the coatings are applied on textiles. An SLR/OLS model was applied to experimental data of biopolymer mixtures to predict the WCA, based on the presence and ratios of coating components. The model predicted the WCA values accurately, proving the potential for guiding the design of multifunctional coatings, by enabling rapid screening and optimisation of the formulations, reducing reliance on extensive laboratory experimentation and consumption of chemicals.

DOI
https://doi.org/10.18690/um.fkkt.1.2026.10

ISBN
978-961-299-130-2

Keywords:
biopolymers, functional materials, coatings, PFAS, data modelling, simple linear regression

University of Maribor Press

1 Introduction

The environmental and health concerns associated with per- and polyfluoroalkyl substances (PFAS) have accelerated the search for sustainable hydrophobic coatings. While PFAS are used widely for imparting hydrophobicity and oleophobicity in textile and packaging materials, they are persistent, bioaccumulative and proven to be toxic. Biopolymers, such as chitosan, cellulose and starch, represent a promising alternative capable of forming functional coatings (Verbič et al. 2025, Golja et al. 2025, Calvo et al. 2024). However, developing and optimising these formulations to achieve the desired functional properties remains difficult, due to complex interactions between the compounds. The traditional experimental approaches rely on repetitive synthesis and testing, which is labour-consuming and resource-intensive and limits the pace of innovation. To overcome these challenges, this research proposes a new approach, based on simple linear regression (SLR), to accelerate and target the development of coating formulations better. By correlating the formulation parameters with functional performance, measured as the water contact angle (WCA), this method enables more efficient and predictive design of hydrophobic coatings.

Recent advances in machine learning (ML) have introduced powerful data-driven workflows for modelling and even inverse design of different formulations in material development (Xie et al., 2025; Wheatle et al., 2020). ML methods can capture complex nonlinear relationships in high-dimensional formulation spaces and complement the classical statistical approaches. In practice, most formulation development systems exhibit nonlinear behaviour due to complex mixtures and interactions between the compounds, which traditionally necessitates the use of more advanced statistical tools. Among these, full quadratic polynomial regression, typically implemented within the framework of response surface methodology (RSM) is applied widely for optimisation. RSM uses regression-based mathematical models to describe how the formulation variables influence the material properties to identify optimal compositions (Chen & Chen, 2025; Elganidi et al., 2022). While these models require only marginally more computational power than SLR, they demand more deliberate experimental design, larger datasets, and more extensive analysis to account for the curvature and interaction effects between the components. However, *in silico* workflows do not always require highly complex models at the earliest stages. In fact, SLR remains a valuable and underexplored tool

for preliminary screening, which can provide a valuable starting point for narrowing down viable formulations.

Despite the fact that SLR models capture only first-order relationships, they are fast, interpretable and effective for preliminary screening. SLR models capture the relationships between variables using a straight line, known as a regression line. This line represents the best fit through the data points (i.e., the best representation of correlation between the variables), and is chosen to maximise the accuracy of model predictions (Miller et al., 2022). Using a regression line, the relationships between the formulation variables (e.g., the presence of different compounds and their ratios) and material responses (e.g., the WCA of the final coated material) can be computed easily. The line of regression is defined by a set of parameters, and there are several methods available for obtaining the estimates of these parameters. One of the methods is ordinary least squares (OLS), arguably the most popular method for estimating the parameters (Su, 2012). It is based on minimising the quantity of the residual sum of squares – residuals being the vertical distances from the fitted line and the measured y-values (Weisberg, 2005). Using this approach, a range of promising candidate formulations can be generated, which can then be refined using more advanced methods, or validated through actual experimental testing. Notably, our literature review revealed that there are no published studies investigating the use of SLR models for prediction of the targeted properties of biopolymer coating formulations, highlighting the need to evaluate the potential of this *in silico* screening followed by targeted experimental verification to support the development of biopolymer blends with desired functional properties in a more efficient manner.

Through data-driven modelling, the correlation between the formulation composition and hydrophobic performance of the coated end-product can be explored systematically, providing valuable insights for optimising coating recipes. The data-driven model enables the predictions of new formulations, based on modified ratios and concentrations of the existing compounds, that have not been yet explored experimentally, along with the corresponding WCA values that these formulations are expected to achieve. This allows researchers to prioritise only the most promising candidates for laboratory validation, thereby reducing the reliance on extensive experimentation. Our study explores whether SLR is an appropriate model, and to what extent strong agreement between the generated predictions and experimentally measured properties is provided, thereby assessing its suitability as a

reliable early-stage screening method that can support resource-efficient and environmentally responsible laboratory development.

2 Materials and methods

2.1 Materials and coating preparation

Six coating compounds were used for the coating formulations: deionised water, alkyl ketene dimer (Melamin d.d., Kočevje, Slovenia), sodium alginate (Sigma-Aldrich, Steinheim, Germany), cellulose nanofibrils (Sappi Valida, Maastricht, Netherlands), corn starch (Sigma-Aldrich, Steinheim, Germany), and agar (Sigma-Aldrich, Steinheim, Germany). Each component was dispersed in water and mixed in a blender. The formulations were prepared by combining these compounds in defined ratios, depending on the specific mixture certain compounds were included, while others were not, resulting in a diverse set of coatings. All the formulations were prepared fresh, and applied immediately to the textile substrates to ensure consistent coating formation. The coated textiles were then dried under controlled laboratory conditions prior to the WCA measurements (Verbič et al., 2025).

2.2 Characterisation of the hydrophobicity

The surface wettability of the coated samples was determined by measuring the WCA values using a tensiometer (Theta T200, Biolab scientific, Sweden). A 3 μL droplet of water was placed on the coated textile surface and the contact angle was measured after 5 seconds. Three measurements were performed on each sample.

2.3 Data preparation, modelling and validation

The dataset consisted of 12 different coating formulations (and an additional uncoated sample) with 11 variables, each described by the relative concentrations of each compound present in the mixture (6 variables) and sample preparation (2 variables), and three measurements of the WCA values of the final coated sample (3 variables), i.e., 312 datapoints.

To develop a predictive model capable of capturing the relationship between the formulation variables and the resulting WCA accurately, an SLR framework was adopted based on the OLS method. Within this framework, the WCA value was treated as the dependent response variable, while the relative percentages of the six components present in the coating formulation were incorporated as explanatory predictors. This setup allowed us to quantify how variations in the formulation composition influenced the observed wettability of the coating surface.

The trained model was then used to generate new, computationally predicted formulations, by sampling within the experimentally feasible ranges of these three input variables. The predicted formulations were then synthesised in the wet-lab following the predicted composition, and evaluated using the same coating and WCA measurement procedures as the initial samples. The newly measured data were incorporated into the dataset for further model refinement.

3 Results and Discussion

To investigate whether an SLR model can support early-stage formulation screening for targeted development reliably, the modelling workflow was structured into four consecutive steps: i) data acquisition from wet-lab experiments, ii) model training, iii) forward projection, and iv) experimental verification with a database update (Figure 1). This combined experimental-computational loop ensured that the model was developed on real experimental inputs, and that its predictions could be validated and improved.

The input data consisted exclusively of the experimentally produced biopolymer coating formulations and their measured WCA values. For every mixture the compositional variables and the corresponding WCA values of the coated samples were inserted into the dataset. The dataset was inspected for inconsistencies and missing entries, and organised to allow meaningful comparison across the formulations. Before fitting the regression model, the independent experimental measurements obtained for each mixture were averaged to yield a single representative value. This averaging procedure reduced the measurement noise, and ensured that the response variable reflected a consistent estimate of the true WCA for each formulation. The model was trained using the dataset of input variables. Its purpose was not only to reproduce the measured results, but to determine whether

a linear relationship could provide accurate WCA predictions for practical use, and to justify SLR as an appropriate screening tool. The model was then fitted to the dataset to identify relationships between the formulations parameters and measured WCA.

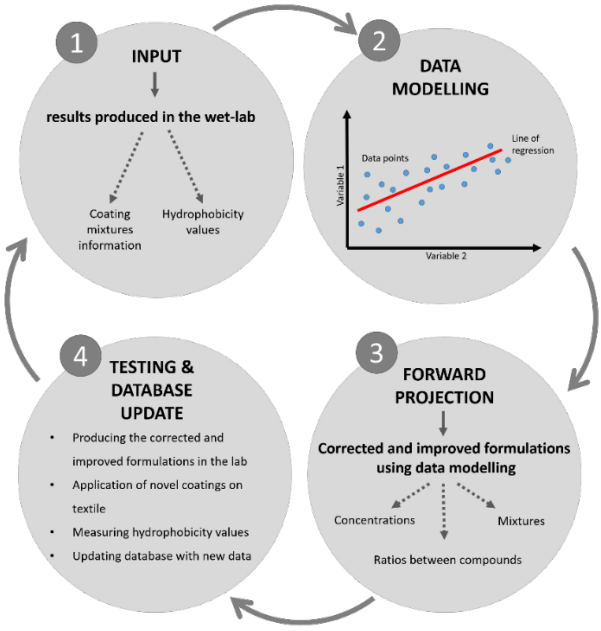

Figure 1: Schematic presentation of the experimental modelling workflow.

The SLR framework based on the OLS method was used to perform the forward projection. The OLS estimator operates by minimising the sum of squared residuals between the observed and predicted values, thereby providing the best linear approximation under the classical assumption that the residual errors are normally distributed with constant variance. This property makes OLS particularly suitable for identifying linear trends in experimental data, and for evaluating the relative importance of the individual formulation variables.

Upon fitting the model, statistical analysis revealed that only three of the six formulation variables (compounds present in the mixture), denoted here as p_1, p_2, p_3, exerted significant effects on the WCA at conventional confidence levels. The remaining variables did not demonstrate statistically meaningful contributions,

suggesting that the wettability of the coating is governed primarily by these three key measurements. The resulting regression model can therefore be expressed in the following general form, Equation (1):

$$WCA = \beta_0 + \beta_1 p_1 + \beta_2 p_2 + \beta_3 p_3 + \epsilon, \quad (1)$$

where WCA is the dependent variable, β_0 represents the intercept term, $\beta_1, \beta_2, \beta_3$ are the estimated regression coefficients associated with the significant formulation variables, and ϵ denotes the residual error term. This model provides a parsimonious yet effective representation of the relationship between formulation composition and surface wettability, and serves as the foundation for the subsequent predictive and experimental validation efforts.

Based on these results, a new dataset was constructed by sampling random compositions systematically within the experimentally feasible ranges defined for the three selected variables. The predicted formulations remained within the chemical space already covered by the experimental work, ensuring practical synthesisability. For each adjusted formulation (i.e. the adjusted ratios between the compounds), the corresponding predicted WCA values using the established model were computed, representing an expected hydrophobic performance. This represents the key results of the modelling stage: the model proposed improved variation of the known formulation and predicted the WCA they should achieve upon testing in the lab. The process allowed to generate a diverse set of formulations that captured a wide spectrum of possible outcomes within the defined parameter space. The resulting dataset not only broadened the scope of the analysis, but also provided a controlled framework for testing the predictive capabilities of the model. In the next stage, these formulations were subjected to experimental validation, enabling us to assess both the accuracy and the robustness of the predictive model rigorously under real-world laboratory conditions. Such validation was crucial to determine whether the model can guide formulation design reliably and ultimately accelerate the discovery of optimised material properties.

To evaluate the predictive performance of the SLR model, a validation step was carried out using formulations newly generated by the model. These formulations were prepared in the wet-lab, according to the predicted compound ratios, applied to textile substrates following the same application protocol used for the initial

samples. The WCA measurements were then performed, to obtain the corresponding experimental hydrophobicity values. The validation process produced two important outcomes. First, the measured WCA values were compared to the model predictions to evaluate the prediction accuracy. The comparison between the predicted and experimentally measured WCA demonstrated that the SLR framework based on the OLS method generated a reliable model in predicting the hydrophobic performance within the range of formulations already explored, confirming its suitability for early-stage screening. In the newly synthesised, model-predicted formulations, the experimentally measured WCA values matched the projected values closely (within 10 %), indicating that the functional properties of the coating were predicted accurately by the model. Notably, several formulations even exceeded their predicted WCA. Second, all the newly obtained data were incorporated into the existing data, to improve the model`s robustness for future iterations and amendments with additional formulation data from other experiments to broaden the exploration of the chemical space.

The results confirm that the SLR framework based on the OLS method can support early-stage screening effectively in the field of Biopolymer-based hydrophobic coatings` development. While the model did not introduce new compounds into the formulation, it predicted new ratios between the existing compounds with increasing WCA successfully, allowing the targeted refinement of formulations without extensive laboratory testing. The ability to predict WCA for adjusted formulations with increased WCA provides a practical decision-making tool for optimising hydrophobic performance in the complex formulation gradients. The predicted WCA values showed a strong agreement with the experimental measurements, establishing SLR/OLS modelling as a valuable tool that can capture dominant trends in small datasets.

4 Conclusions

This study demonstrates that data-driven modelling provides an effective and sustainable approach for developing biopolymer-based hydrophobic coatings as alternatives to PFAS. By applying the SLR framework based on the OLS method, the correlation between coating composition and WCA was modelled successfully, allowing reliable prediction of the hydrophobic performance with increased WCA values. Despite being performed on a relatively small dataset using fewer than 350

data points, the model achieved high predictive accuracy and identified an improved variant of the existing formulations successfully, which were validated experimentally. The integration of data modelling into the formulation development supported rapid screening and optimisation of the coating composition, reducing the experimental workload, material consumption and environmental impact significantly, while accelerating the pace of innovation. Future work will focus on expanding the dataset, exploring other models and incorporating additional variables, which will improve the predictive performance and broaden the model's applicability further.

Acknowledgment

This research was supported by the Horizon Europe project PROPLANET (Grant agreement number 10109842) and the Slovenian Research Agency (Research Core Funding No. P2-0152).

References

Calvo O., Blanes M., Sirvent E.A., Pastor Climent B., Verbič A., Stres B., Golja B., Novak U., Likozar B. (2024). Desarrollo de acabados textiles libres de PFAS mediante alternativas biobasadas. Proceedings of XI Congreso I+D+i, Campus d'Alcoy : Creando sinergias, pp.365-370.

Chen, H. Y., & Chen, C. (2025). Importance of Using Modern Regression Analysis for Response Surface Models in Science and Technology. *Applied Sciences*, *15*(13), 7206.

Elganidi, I., Elarbe, B., Ridzuan, N., & Abdullah, N. (2022). Optimisation of reaction parameters for a novel polymeric additives as flow improvers of crude oil using response surface methodology. *Journal of Petroleum Exploration and Production Technology*, *12*(2), 437-449.

Golja B., Stres B., Novak U., Likozar B., Verbič A. (2025). Eco-friendly hydrophobic textiles: a shift from PFAS to sustainable bio-polymers. Proceedings of crossing boundaries: 50th International Symposium on Novelties in Textiles. p.123-129.

Miller, A., Panneerselvam, J., & Liu, L. (2022). A review of regression and classification techniques for analysis of common and rare variants and gene-environmental factors. *Neurocomputing*, *489*, 466-485.

Su, X., Yan, X., and Tsai, C. (2012). Linear regression. *Wiley Interdisciplinary Reviews Computational Statistics*, *4*, 275–94.

Verbič A., Stres B., Jerman I., Golja B., Žagar E., Martinović V., Logar P., Lavrič G., Prašnikar A., Likozar B., Novak U., Oberlintner, A. (2025). Breaking free from PFAS: biocompatible, durable and high-performance octenyl succinic anhydride (OSA)-modified starch/chitosan coating with ZnO for textile applications. *Carbohydrate Polymers*, 123792.

Weisberg, S. (2005). Applied linear regression. *John Wiley & Sons, 528,* 21-22.

Wheatle, B. K., Fuentes, E. F., Lynd, N. A., & Ganesan, V. (2020). Design of polymer blend electrolytes through a machine learning approach. *Macromolecules*, *53*(21), 9449-9459.

Xie, C., Qiu, H., Liu, L., You, Y., Li, H., Li, Y., ... & An, L. (2025). Machine learning approaches in polymer science: Progress and fundamental for a new paradigm. *SmartMat*, *6*(1), e1320.