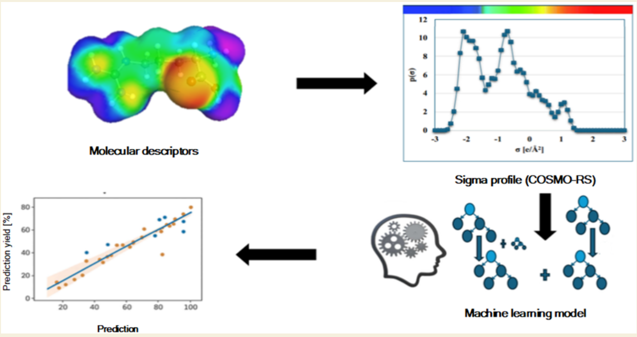
ACS ENGINEERING Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Screening of Protic Ionic Liquids with Catalytic Potential in the Transesterification Reaction Using COSMO-RS and Machine Learning

Luis Alberto Gallo-García, Pedro J. Carvalho, Maria Isabel da Silva Nunes, Nian Vieira Freire, Alessandro Cazonatto Galvao, Daniela Helena Pelegrine Guimarães, and Pedro Felipe Arce*

Cite This: https://doi.org/10.1021/acsengineeringau.5c00098 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Global population growth has led to the use of fossil fuels and global pollution problems. Biodiesel, a renewable and environmentally friendly alternative to petroleum fuel, is produced from organic oils and animal fats, causing food safety issues. Unprocessed crude oils are inexpensive raw materials with a high content of free fatty acids. Ionic liquids (ILs) are used as catalysts for biodiesel production to solve the problems of traditional catalysts. This manuscript proposes the COSMO-RS model and machine learning as predictive tools for screening ILs as catalysts for fatty acid methyl esters (FAME) synthesis. COSMO-RS activity coefficient model was used to obtain the ILs sigma profile and interaction energies (electrostatic-misfit (Emisfit), hydrogen bond (E_{HB}), and van der Waals (E_{vdW})) to correlate the yield of reaction. The machine learning models, such as K-nearest neighbor, Random Forest Regressor, Decision Tree Regressor model, Gradient Boosting Regressor, and Multilayer Perceptrons model, were applied to correlate the above-mentioned properties. The Gradient Boosting Regressor model, using the analysis of the anion and cation sigma profiles, proved to be more efficient than the same model, using the approach of the interaction energies. Based on the screening study, the ILs [L-arginine][Acetate], [L-arginine][HSO_3], and [L-arginine][NO_3] were selected, synthesized, and characterized.

KEYWORDS: *ionic liquids, biodiesel, interaction energy, gradient boosting regressor, jupyter notebook, sigma profile*

1. INTRODUCTION

Clean and renewable energy sources, such as biodiesel, play an important role in mitigating the global challenge, the urgency of combating climate change, and achieving the 2030 Agenda for Sustainable Development. Rising emissions of various greenhouse gases into the atmosphere have serious consequences for our planet; therefore, the transition to renewable energy is more crucial than ever. Biodiesel, obtained from renewable resources such as vegetable oils, recycled oils, and animal fats, can help reduce gas emissions and mitigate climate change to ensure a sustainable future for our planet.[1,2]

The traditional methods used to produce biodiesel from fats and oils are transesterification, pyrolysis, and emulsification. Transesterification is a reaction commonly used in industry because it produces an environmentally friendly biofuel that is highly compatible with the currently used diesel engines. It involves fats and oils in the presence of alcohol and an effective catalyst, generating alkyl esters and glycerol as a bioproduct.[3]

Biodiesel production involves the use of different types of catalysts for the transesterification of triglycerides. These include homogeneous and heterogeneous catalysts, as well as biocatalysts such as enzymes (enzymatic catalysis).[4]

Homogeneous base catalysts include metal oxides and alkaline liquids such as potassium methoxide, sodium methoxide, carbonates, sodium hydroxide, and barium hydroxide.[5] Homogeneous acid-catalyzed transesterification offers an advantage over homogeneous base-catalyzed transesterification because the presence of free fatty acids does not deactivate the acid catalyst, and both esterification and transesterification can be catalyzed simultaneously, although it requires long times and high temperatures.[6] However, when it comes to esterification and transesterification reactions, the use of a homogeneous catalyst causes many problems when

Received: November 3, 2025
Revised: January 7, 2026
Accepted: January 8, 2026

sulfuric acid is used in esterification, causing corrosion of the reactor.

Compared to homogeneous catalysts, heterogeneous catalysts offer several benefits, including easy regeneration and lower corrosiveness and are more environmentally friendly. However, they have some disadvantages, such as the use of high temperatures and a high alcohol molar ratio.[7]

To this end, the synthesis of more environmentally friendly and sustainable catalysts for biodiesel production is urgently needed. It is necessary to explore new, efficient catalysts to help solve or mitigate the aforementioned problems.[8] Ionic liquids emerge as alternative catalysts that can meet these needs, given their unique physicochemical properties, such as thermal stability, nonflammability, nonvolatility, immiscibility with organic solvents, and their recoverability.[9] Basic ionic liquids such as cholinium arginate [Cho][Arg], Tetrabutylammonium arginine [TBA][Arg] have been used as substitutes for conventional catalysts for the transesterification of vegetable oils and the acidic ionic liquids, such as 1-butylsulfone te-3-methyl imidazolium hydrogen sulfate ([BSO_3Hmim][HSO_4]) and 1-(4-sulfonic acid) butylpyridinium hydrogen sulfate ([HSO_3−BPyr][HSO_4]), have been used for the esterification of fatty acids or animal fats due to their high catalytic activity.[10−13] As observed in the literature reviews by O'Connor et al.[14] and Zhang and Sun,[15] it is notable that there is catalytic potential with ionic liquids for biodiesel synthesis; however, there are a high number of ionic liquids as potential candidates, insufficient characterization and scarcity of research in this area.

COSMO-RS is a thermodynamic model that predicts the properties of pure fluids or mixtures based on their chemical structure using a quantum chemical approach. The model employs a statistical thermodynamic approach and the "COnductor-like Screening MOdel" (COSMO) for efficient continuous dielectric solvation calculations. It represents molecules as surface segments, and the chemical potential is calculated from the interaction energies. The total chemical potential is the sum of the segment contributions. Studies have linked the polarity of ionic liquids to quantum chemical parameters, demonstrating their effectiveness as a correlation tool.[16−19]

On the other hand, machine learning (ML) is a data-driven modeling technique classified into several types. Supervised learning uses labeled data for specific tasks, while unsupervised learning seeks patterns in unlabeled data. In the last two decades, ML has become a widely used technology in commercial settings within artificial intelligence (AI). Many AI developers find it easier to train systems using examples of desired input−output behavior rather than manually programming responses for all possible inputs.[20,21]

Furthermore, the polarity of the molecules stored in the sigma profile of ionic liquids obtained with the COSMO-RS model has not been fully explored with machine learning models to predict the catalytic capacity of ILs that has not yet been characterized. Therefore, this study aimed to predict the catalytic capacity of ionic liquids that have not been used in catalysis for the synthesis of fatty acid methyl esters (FAMEs). To this end, COSMO-RS was used as a predictive tool to generate the sigma profile of ionic liquids and use them as molecular descriptors in machine learning models with the aim of correlating the described properties.

1.1. COSMO-RS Activity Coefficient Model

COSMO-RS is a quantum chemistry-based method for predicting thermodynamic properties of liquids and solutions where molecules are represented as a collection of interacting surface segments, each characterized by its screening charge density (σ). COSMO-RS calculations are performed for all molecules involved, generating a 3D polarization density distribution on the molecular surface. Then, the 3D polarization density is converted into a distribution function called the σ-profile ($p(\sigma)$), which describes the polarity of each surface segment. The chemical potential of a molecule X in solvent S is then calculated by integrating all its surface segments and adding combinatorial and dispersive contributions. This approach allows the COSMO-RS to predict various thermodynamic properties, including activity coefficients, solubility, partition coefficients, vapor pressure, and free energy of solvation. The method's strength lies in its ability to combine quantum chemical accuracy with statistical thermodynamics, enabling predictions for a wide range of systems without the need for system-specific parameters.

From the molecular sigma profiles, the sigma profiles of the whole system/mixture (S) can be derived as the sum of the mole fraction of the sigma profiles of the components weighted with their mole fraction in the mixture x:

$$ps(\sigma) = \frac{\sum_i x_i p^{X_i}(\sigma)}{\sum_i x_i A^{X_i}} \tag{1}$$

$$E_{\text{misfit}}(\sigma,\ \sigma') = a_{\text{eff}} \frac{\sigma'}{2} (\sigma + \sigma')^2 \tag{2}$$

$$E_{\text{HB}} = a_{\text{eff}} c_{\text{HB}} \min(0;\ \min(0;\ \sigma_{\text{donor}} + \sigma_{\text{HB}}) \\ \max(0;\ \sigma_{\text{acceptor}} - \sigma_{\text{HB}})) \tag{3}$$

$$E_{\text{vdW}} = a_{\text{eff}(\tau_{\text{vdW}} + \tau'_{\text{vdW}})} \tag{4}$$

In this context, the interaction energy $E_{\text{INT}}(\sigma,\sigma')$ includes electrostatic interactions (misfit energy, E_{misfit}), hydrogen bonding (E_{HB}), and van der Waals (E_{vdw}):

$$E_{\text{INT}}(\sigma,\ e,\ \sigma',\ e') \\ = \frac{E_{\text{MF}}(\sigma,\ \sigma') + E_{\text{HB}}(\sigma,\ \sigma') + E_{\text{vdW}}(e,\ e')}{a_{\text{eff}}} \tag{5}$$

where e and e' are the contacting surface segments.[22,23]

$$\mu_{\text{S}}(\sigma) = -\frac{RT}{a_{\text{eff}}} \ln\left[\int ps(\sigma') \exp\left(\frac{1}{RT}(a_{\text{eff}}\ \mu_{\text{S}}(\sigma') \\ - E_{\text{misfit}}(\sigma,\ \sigma') - E_{\text{HB}}(\sigma,\ \sigma'))\right) d\sigma'\right] \tag{6}$$

$\mu_{\text{S}}(\sigma)$ is called the potential σ and can be interpreted as the affinity of solvent S for the surface of polarity σ. Finally, the pseudochemical potential of compound X in system S can be calculated by integrating $\mu_{\text{S}}(\sigma)$ over the surface of the compound, eq 7.[23]

$$\mu_{\text{S}}^{X} = \mu_{\text{C,S}}^{X} + \int p^{X}(\sigma) \mu_{\text{S}}(\sigma) d\sigma \tag{7}$$

$\mu_{\text{C,S}}^{X}$ is a combinatorial term commonly used in chemical engineering models.

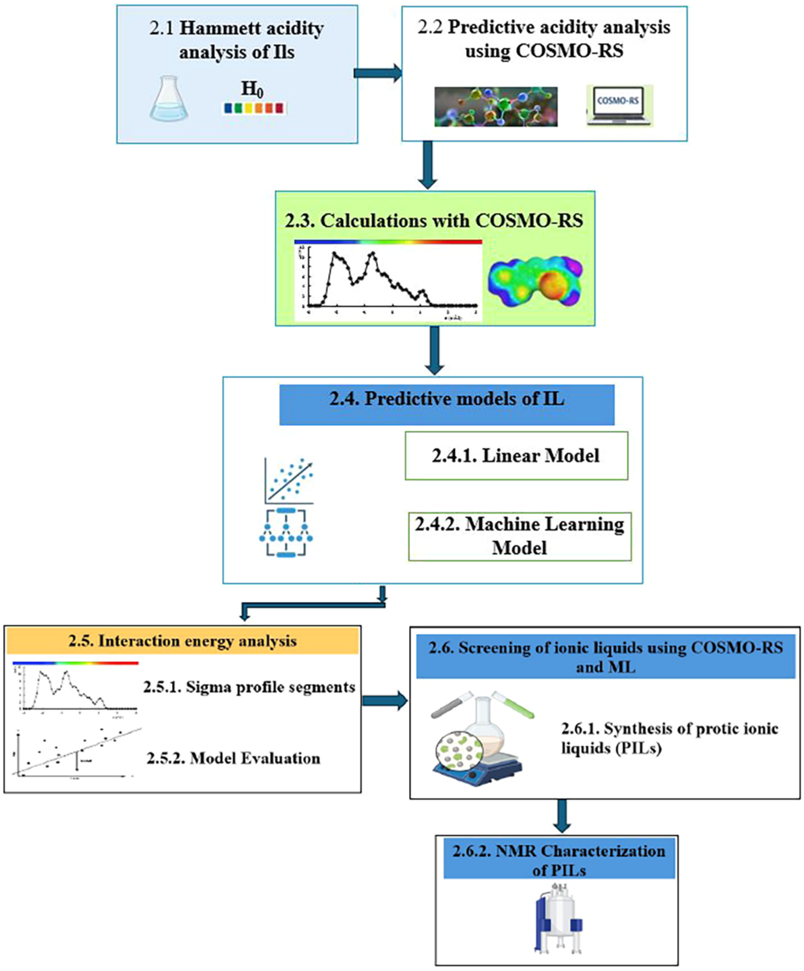

2.1 Hammett acidity analysis of Ils
H_0
2.2 Predictive acidity analysis using COSMO-RS
COSMO-RS
2.3. Calculations with COSMO-RS
2.4. Predictive models of IL
2.4.1. Linear Model
2.4.2. Machine Learning Model
2.5. Interaction energy analysis
2.5.1. Sigma profile segments
2.5.2. Model Evaluation
2.6. Screening of ionic liquids using COSMO-RS and ML
2.6.1. Synthesis of protic ionic liquids (PILs)
2.6.2. NMR Characterization of PILs

Figure 1. Workflow of the methodology used.

2. MATERIALS AND METHODS

The overall workflow is described in Figure 1, and the steps are indicated throughout the explanation. In this study, the Hammett acidity of ionic liquids (ILs) with catalytic capacity reported in the literature was initially analyzed (Supporting Information, Table S1). Then, predictive acidity analysis was performed using COSMO-RS, utilizing electrostatic misfit (E_{MF}), hydrogen bond (E_{HB}), and van der Waals (E_{vdW}) energies (Supporting Information, Table S2). Consequently, acidity values were analyzed against yield, considering the influence of the cation and anion with the aim of demonstrating any correlation. Subsequently, machine learning models were used, employing interaction energies, and descriptors based on sigma profiles, specifically the areas under the sigma profile (S-profile) (Tables S1−S5), were analyzed to correlate five segments of areas below the S-profile curve of the ionic liquids (Supporting Information, Tables S4 and S5). Analysis of the area under the curve segments, using statistical criteria such as the coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (RMSE), showed that the Gradient Boosting Regressor (GBR) model achieved higher statistical resolution in predicting FAME. Therefore, based on these FAME performance criteria, ionic liquids with potential catalytic characteristics were selected and identified for use in FAME synthesis. These ionic liquids were characterized by ^{1}H and ^{13}C NMR.

2.1. Qualitative Analysis of the Hammett Acidity (H_0) of Ionic Liquids

Many chemical transformations are sensitive to the presence of protons. Therefore, knowing the pK_a value is essential for deciding on their possible applications as reaction media. A common and effective method for assessing the acidity of Brönsted acids was the Hammett method, based on the Hammett acidity function (H_0), in which a basic indicator was used to trap the acidic proton.[24]

The Brönsted acid properties of a substance are based on the Brönsted-Lowry theory. According to this theory, a Brönsted acid and base are defined as substances that donate or accept a hydrogen ion (H^+) or a proton, respectively. Therefore, a Brönsted acid ionic liquid (BAIL) can be defined as an ionic liquid that can donate a hydrogen ion (H^+) or a proton. Acidic ionic liquids are prepared by reacting a Brönsted base with a Brönsted acid. BAILs with one or more acidic hydrogens residing on N or O atoms are also known as protic ionic liquids (PILs). There are attempts to qualitatively correlate the calculated H_0 values of BAILs with catalysis activities in chemical reactions such as transesterification. These experiments show that BAILs containing CF_3SO_3 and HSO_4 anions with lower H_0 values generate higher conversions and yields.[25]

Data were collected from the literature on various ionic liquids that have demonstrated catalytic capacity to analyze whether the acidity of the IL could significantly impact the esterification/transesterification reaction.[12,26] After compiling information on the acidity of different ionic liquids with catalytic capacity, it was initially plotted H_0 against FAME yield (%) graph (Supporting Information, Table S1 and Figure S1). Performance data on FAME and H_0 were obtained from the literature.

2.2. Predictive Acidity Analysis Using COSMO-RS

The method proposed by Kurnia et al.[27] was used as a predictive tool to determine the acidity of the hydrogen bond of the interaction energies of ILs, as presented in eq 8.

$$\alpha = -(0.0164 \pm 0.0073)E_{\text{misfit}} + (0.0474 \pm 0.0046)E_{\text{HB}} + (0.0017 \pm 0.0014)E_{\text{vdW}} + (0.6934 \pm 0.1696) \quad (8)$$

Table S1 presents calculated values for the acidity of the ionic liquids. After optimizing the structures of the ILs, the acidity of the ILs was determined. Consequently, the acidity versus yield values, as illustrated in Figures S2 and S3 (Supporting Information), were analyzed by considering the influence of the cation and anion to demonstrate a potential correlation.

2.3. Calculations with COSMO-RS

The structures of the ionic liquids were generated using ChemDraw 22.2.0 and Chem3D 22.2.0 software. In this context, using the Turbomole program package version 4.5 (TmoleX19 interphase), the molecular geometry of the ionic liquids was predicted and saved, along with the charge density, charge distribution, and molecular surface, in the COSMO format. The sigma profiles used in this work were generated from the COSMO-RS thermodynamic model. Calculations were performed using the COSMOtherm software with the BP_TZVP_21_.ctd parametrization, utilizing the ".cosmo" files generated after optimizing the geometry (COSMOtherm version 21, COSMOlogic GmbH & Co KG. Leverkusen, Germany). Separate files were used for the IL cations and anions, using an equimolar cation−anion mixture of 65 IL to determine the interaction energies (E_{misfit}, E_{HB}, and E_{vdw}) of a pure IL, with the COSMOtherm software (version 21, BP_TZVP_21_.ctd parametrization) using the molecular surface charge density parameter file (Supporting Information, Table S2). After optimizing the structures of the ILs (Supporting Information, Table S3), the acidity of the ILs was determined.

2.4. Predictive Models of IL

2.4.1. Linear Model. This stage involved selecting the input variables, which are the independent variables of the model. In this regard, two models were proposed. The first model considered the interaction energies of 65 ionic liquids (E_{misfit}, E_{HB}, and E_{vdw}) with documented catalytic capacity from the literature as presented in Table S2 of the Supporting Information.

The linear regression model was applied by using the Linear Regression function from the scikit-learn library. The dependent variable was the yield (FAME). The independent variables included the interaction energies E_{misfit}, E_{HB}, and E_{vdw} associated with the cation (E_{misfitC}, E_{HBC}, and E_{vdwC}) and the anion (E_{misfitA}, E_{HBA}, and E_{vdwA}) of 65 ionic liquids.

Data visualization was performed using the Python Seaborn library. The Pearson correlation was applied to check whether there was a correlation between the independent variables and the dependent variable, using the *corr()* function, and visualized through the *heatmap()* function of the Seaborn library, as shown in eq 9.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}} \quad (9)$$

In multiple regression analysis, the term "multicollinearity" refers to a linear relationship among the independent variables. This occurs when regression models include variables that are significantly correlated with not only the dependent variable. The presence of multicollinearity increases the variance of the regression coefficients, making them unstable, which creates challenges when the coefficients.

To determine whether multicollinearity exists among the independent variables, the Variance Inflation Factor (*VIF*) is employed for this purpose, utilizing eq 10.

$$VIF = \frac{1}{1 - R^2} = \frac{1}{\text{tolerance}} \quad (10)$$

Tolerance is the inverse of the *VIF*. The lower the tolerance, the more likely there is multicollinearity among the variables. A *VIF* value of 1 indicates that the independent variables are uncorrelated. If the *VIF* value is between 1 and 5, it suggests that the variables are moderately correlated. The *VIF* value range to pay most attention to is that between 5 < *VIF* < 10, as this indicates that the variables are highly correlated.[28]

2.4.2. Machine Learning Models. The computing platform used in this work was Jupyter Notebook, an open-source web environment for interactive computing that supports various programming languages, such as Python, *R*, LaTeX, JavaScript, and many others.[29] The Pandas library was utilized to import and manipulate data, tables, and data frames. NumPy, a Python library for processing arrays, was used to manage the data.

The *train_test_split* function from the scikit-learn data science library was used to separate the data into training and test sets. All models were trained using 80% of the data, while the remaining 20% was reserved for testing. The data were normalized and standardized with StandardScaler, a preprocessing technique provided by scikit-learn. For the DTR model, hyperparameters were optimized using the GridSearchCV approach, which helps identify overfitting. These hyperparameters allow to find the ideal combination to optimize the model's performance, making it more accurate and robust.

Using the Jupyter Notebook with Python programming language, all input variables were utilized to predict FAME performance through machine learning algorithms. The primary focus is on the accuracy of the models, as well as the prediction error. Analyzing errors aids in understanding the bias and variance of the model. The error rate is commonly termed bias, with the selection of input data being the most influential parameter. The variance of a model denotes the decrease in accuracy when assessing the model's performance on test data compared to training data.[30]

In the final model, the number of nearest neighbors for the KNN model was set to three. Additionally, the model conducted an intensive force search to identify the closest locations, and the distance weight function was applied in the prediction process. In the context of the algorithm being analyzed, "distance" describes how weight is assigned to points, based on the inverse of their distance. This implies that points closer to a query point will have a greater influence than those further away.[31] In summary, the KNN model criteria used are *n_neighbors = 2, weights = "distance", algorithm = "brute"*, and *leaf_size = 2.*

The Gradient Boosting Regressor (GBR) algorithm is frequently employed to predict energy consumption because of its high accuracy.[32] However, achieving an accurate prediction of the FAME yield in this study necessitates a proper adjustment of the GBR parameters, which demands considerable time.

For the Gradient Boosting Regressor model, the criteria used are *n_estimators = 100, max_depth= 4, max_features = "log2", min_samples_leaf= 1, min_samples_split= 2, criterion= "squared_error", learning_rate= 0.1,* and *tol = 0.0001.*

For the Random Forest Regressor model, the criteria used were *n_estimators = 100, random_state = 42, criterion = "squared_error", max_depth = 10, max_features = 1,* and *bootstrap = False.* For the MLPRegressor model, the parameters used are *hidden_layer_sizes = (100), max_iter = 20,000, activation = "logistic", learning_rate = "invscaling", tol = 0.0001, α = 0.0001,* and *solver = "lbfgs"*. The solver for weight optimization in the boosted MLP was chosen to be *"lbfgs"*, which is an optimizer that falls into the category of quasi-Newton methods.

The Decision Tree Regressor model utilized the following criteria: *splitter = "random", max_depth = 80, min_samples_split = 3, min_samples_leaf = 1, max_features = "log2", criterion = "absolute_error", random_state = 42.*

2.5. Analysis of the Interaction Energies of Ionic Liquids

Machine learning models have been proposed, including K-Nearest Neighbor (KNN), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), Decision Tree Regressor (DTR), and the multilayer perceptron (MLP) neural network model. These models were developed to analyze the interaction energies of the cation and anion alongside the yield (% FAME) of the catalytic activity of 65 ionic liquids employed by researchers for the esterification/transesterification of various oils. The effectiveness and accuracy of each model were assessed based on statistical factors, including the correlation coefficient (R^2), mean absolute error (MAE), and root-mean-square error (RMSE). The results were compared, and the model demonstrating the best performance according to the statistical criteria was selected.

2.5.1. Analysis of Sigma Profile Surface Area Segments (Molecular Descriptors). Sigma profiles are molecular descriptors that represent the polarity of molecules and, consequently, their capacity to form and engage in intermolecular interactions such as hydrogen bonds. A sigma profile is a probability distribution of the surface charge density of a molecule or mixture.[33,34]

Torrecilla et al.[35] demonstrated how the σ profile qualitatively describes the different electronic nature of cations and anions. The authors showed that the charge distribution area beneath the σ profile can serve as a suitable molecular descriptor of solvent properties, with the significant advantage of being a parameter derived from quantum chemistry defined within a limited polarity scale (±0.025 e/Å^2 for most compounds). These researchers, considering the 61 levels of charge distribution $p^X(\sigma)$, defined the σ-profile of ionic species in ILs within the range of ±0.03 e/Å^2, producing 61 values of the $S\sigma$ profile. The work contributed to the development of a neural network model for predicting the toxicological effect of ILs on a rat leukemia cell line (Log EC$_{50}$ IPC-81) across a diverse range of compounds including imidazolium, pyridinium, ammonium, phosphonium, pyrrolidinium, and quinolinium ILs.

In the work of Alkhatib et al.,[36] the σ profiles generated were divided into eight regions, each featuring a step size of 0.00625 e/Å^2. These regions were utilized to calculate the molecular descriptors, specifically the $S\sigma$ profile, as integrals of the area under the curves of the σ profile across these eight regions. As noted, researchers have also segmented the sigma profile into several parts to provide a more comprehensive description of the hydrogen bond interactions. Moreover, Hsieh et al.[37] proposed separating the sigma profile into a non-hydrogen bonded hydroxyl group and a nonhydroxyl group.

In this work, the sigma profiles of 65 ILs were obtained and categorized into five distinct regions within the range of −3.00 to +3.00 with the intervals of approximately 1.2 e/Å^2. The surface shielding charge area of each region is obtained through integration, similar to how the σ profile is imported into Aspen Plus software by using the COSMO-SAC thermodynamic model.

Two approaches were employed to analyze the areas under the curve of the ILs concerning the yield (FAME). In the first approach, the areas of the cation and anion of each IL were summed, resulting in a total of five areas representing the combined cation and anion regions (S_{1CA}, S_{2CA}, S_{3CA}, S_{4CA}, and S_{5CA}) as illustrated in Figure 2. The calculated values for each area are presented in Table S4 (Supporting Information). In the second approach, each area of the cation and anion was utilized in the model, totaling five areas of the cation and five areas of the anion of the IL (S_{1C}, S_{2C}, S_{3C}, S_{4C}, S_{5C}, S_{1A}, S_{2A}, S_{3A}, S_{4A}, and S_{5A}) as presented in Figures 3 and 4. Table S5 shows the calculated values for the areas of the cation and anion. The experimental data set was randomly divided into training (80%) and test (20%) subgroups.

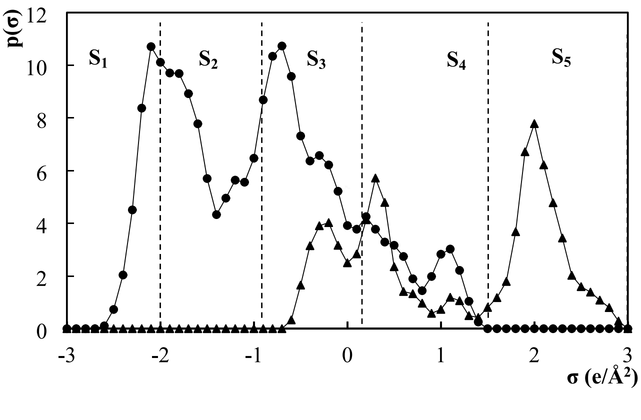


Figure 2. Area segments of the model sigma profile of the ionic liquid [L-Arginine][Acetate], S_{CA} approach: ● L-arginine, ▲ acetate, continuous line is just for orientation.

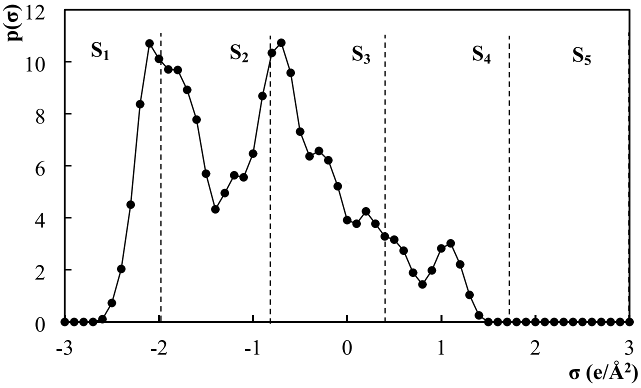


Figure 3. Area segments of the model sigma profile of the cation [L-arginine], $S_{C\text{-}A}$ approach.

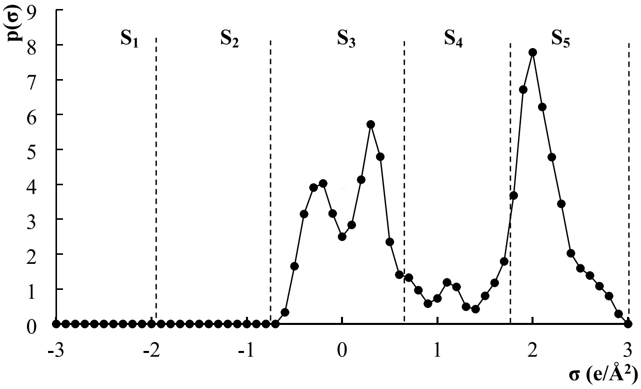


Figure 4. Area segments of the model sigma profile of the anion [Acetate], $S_{C\text{-}A}$ approach.

To calculate the area under the curve, the *Scipy.integrate* function utilizing the Simpson integration method was employed. Following a comprehensive evaluation of the statistical and predictive capabilities of the ML model, the pickle format/module was used to serialize the model into a file for future access. The *dump()* function was applied to write a Python object to a file for subsequent use. The *Joblib* library was utilized to save the trained model, which was employed to determine the yield (% FAME) of the ILs of interest by importing the areas under the sigma profile curve of these ILs as input data. A total of 298 ILs were analyzed, primarily focusing on those based on amino acids, ammonium, acids, and imidazolium, which featured cations and anions relevant to the research objectives.

2.5.2. Criteria for Evaluating Models. The effectiveness, performance, and accuracy of the model were analyzed according to statistical factors, such as the coefficient of determination (R^2), the mean absolute error (MAE), and the root-mean-square error (RMSE). In the following equations, according to Chicco et al.,[38] x_i is the i^{th} predicted value, and the element y_i is the i^{th} actual value. The regression method predicts the x_i element for the corresponding y_i element of the ground truth data set. It defines the constant $\bar{y}$ as the average of the true values, eq 11.

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=n}^{n} (x_i - y_i)^2}{\sum_{i=n}^{m} (\bar{y} - y_i)^2} \tag{12}$$

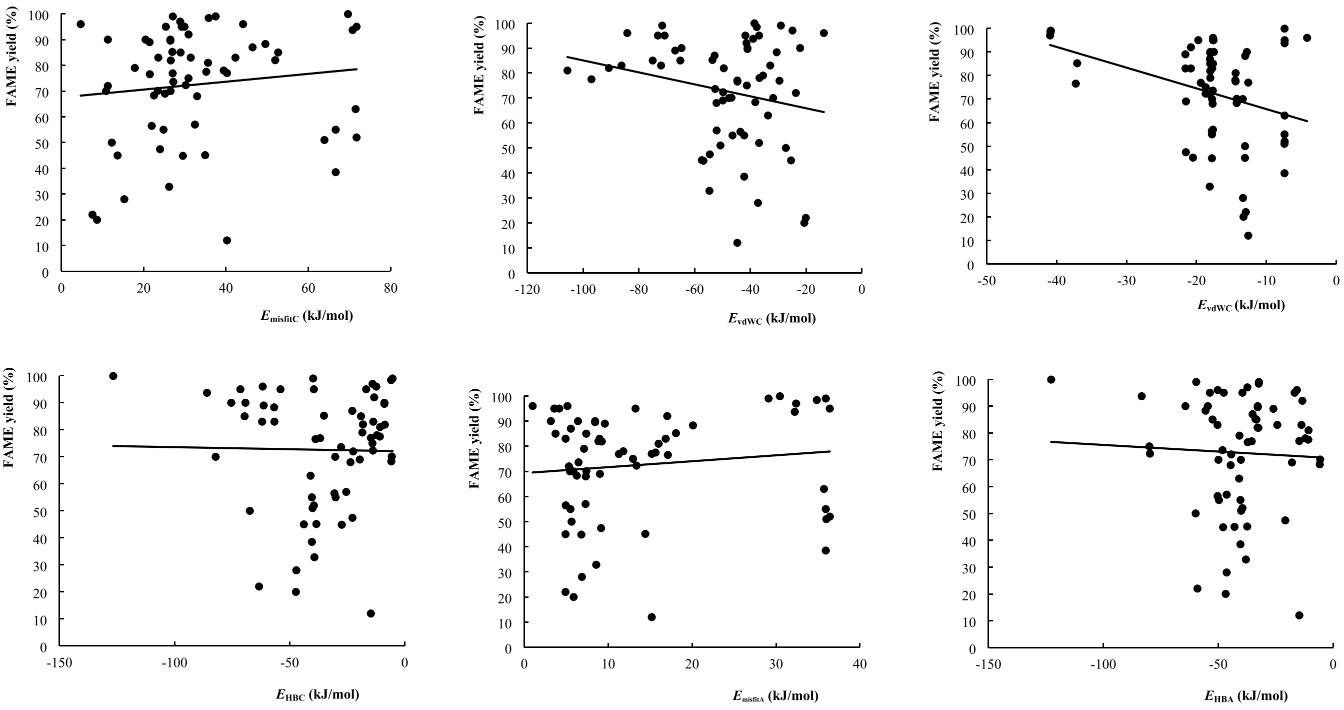


Figure 5. Interaction energies of ionic liquids vs FAME yield.

The coefficient of determination (R^2) defined by Wright in 1921 can be interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variables, eq 12.[38]

The mean absolute error (MAE) represents the average of the absolute values of the errors that indicate the deviation from the true probability. This is mathematically expressed in eq 13.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n} \tag{13}$$

On the other hand, the root mean squared error (RMSE) is a popular performance evaluation metric for models because it is interpretable as the standard deviation of prediction errors. This is written by Otchere et al.[30] as is shown in eq 14.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - y_i)^2}{n}} \tag{14}$$

2.6. Screening of Ionic Liquids Using COSMO-RS and ML

Considering the FAME yield predictions from the ML model were considered, other parameters were analyzed, including the chemical structures of the cation and anion, the cost of the reagents, their availability in the laboratory, and the difficulty of the synthesis. Based on these criteria, three ILs were selected: [L-arginine][Acetate] (Figure S4), [L-arginine][NO_3] (Figure S5), and [L-arginine][HSO_3] (Figure S6) to be synthesized and characterized by ^1H and ^{13}C NMR (Supporting Information).

2.6.1. Synthesis of Protic Ionic Liquids (PILs). The synthesis of amino acid–based PILs was achieved using the Brønsted acid–base neutralization method proposed by Sharma et al.[39] and Martins et al.,[40] with a molar ratio of 1:1.2 (cation:anion). In summary, the aqueous solutions of L-arginine were placed in a three-neck round-bottom flask (in an ice bath) under stirring, equipped with a condenser at 12 °C and a drip funnel at the center of the flask. The acid was added drop by drop using the dropping funnel at approximately 283.15 K until all of the acid was added. After the acid was added, the reaction mixture was stirred for 24 h at room temperature under a nitrogen atmosphere. Once the reaction was complete, the mixture was evaporated at 60 °C and 50 mbar for 5 h to concentrate the IL. Finally, the reaction mixture was dried for 36 h at 323.15 K using a high vacuum pump. Figure S4 of the Supporting Information illustrates the experimental setup.

2.6.2. Characterization of PILs by ^1H and ^{13}C Nuclear Magnetic Resonance. The synthesized PILs were characterized by using NMR spectral techniques (^1H and ^{13}C) on a Bruker Avance III 300 MHz spectrometer with 5 mm glass tubes. A total of 50 mg of sample was used in 500 μL of deuterium oxide ($D_2O/2H_2O$) as the solvent. A capillary containing a solution of sodium fluoride (NaF), 4 mg in 5 mL of D_2O, was placed in each glass tube and served as an external reference.

3. RESULTS AND DISCUSSION

3.1. Analysis of the Hammett Acidity of Ionic Liquids with Catalytic Capacity

The literature indicates that Hammett acidity (H_0) is the primary parameter that most researchers agree accounts for the catalytic activity of a family of ionic liquids, as illustrated in Figure S1 of the Supporting Information.

The research by Gao et al.[26] demonstrates that the IL [Im(N(CH_2)$_3$$SO_3$H)$_2$][$HSO_4$], which has a H_0 of 0.85, exhibits superior catalytic performance with an 86% FAME yield compared to other IL with higher acidity that were studied. Similarly, Masri et al.[41] reported enhanced catalytic properties with the IL [TMEDADBS][HSO_4]$_2$, which has an H_0 of 2.370 and an FAME yield of 59%.

Tankov et al.[42] synthesized protic IL, demonstrating that the synthesis of FAME was influenced by acidity, with yields of 53% using the catalyst pyridinium hydrogen sulfate (H_0 of 1.62), pyridine nitrate (H_0 = 1.84) yielding 12%, and 4-amino-1*H*-1,2,4-triazolium nitrate (H_0 = 2.56) yielding 5%. Li et al.[43] developed functionalized IL based on ethanediamine (EDA), diethylenetriamine (DETA), triethylenetetramine (TETA), and tetraethylenepentamine (TEPA). The H_0 results were 0.971 for [EDA-PS][HSO_4], 0.981 for [DETA-PS][HSO_4], 0.980 for [TETA-PS][HSO_4], and 0.992 for [TEPA-PS]-[HSO_4]. The FAME yield ranked as follows: [EDA-PS]-[HSO_4] (41%) > [DETA-PS][HSO_4] (39%) > [TETA-

Table 1. Cation and Anion Interaction Energies of Ionic Liquids

	$E_{misfitC}$	E_{HBC}	E_{vdWC}	$E_{misfitA}$	E_{HBA}	E_{vdWA}	FAME yield
$E_{misfitC}$	1.00000	−0.12476	−0.19411	0.731978	−0.16901	0.291424	0.120288
E_{HBC}	−0.12476	1.00000	−0.02261	−0.00811	0.673807	−0.28261	−0.01722
E_{vdWC}	−0.19411	−0.02261	1.00000	0.13886	−0.16338	0.11418	−0.21125
$E_{misfitA}$	0.731978	−0.00811	0.13886	1.00000	−0.10285	−0.09846	0.117047
E_{HBA}	−0.16901	0.673807	−0.16338	−0.10285	1.00000	−0.11586	−0.04583
E_{vdWA}	0.291424	−0.28261	0.11418	−0.09846	−0.11586	1.00000	−0.33477
FAME yield	0.120288	−0.01722	−0.21125	0.117047	−0.04583	−0.33477	1.00000

PS][HSO_4] (36%) > [TEPA-PS][HSO_4] (34%). The authors concluded that Hammett's acidity was the main parameter affecting catalysis.

Moreover, Fan et al.[12] analyzed the influence of the [HSO_4]$^-$ anion on the synthesis of FAME from oleic acid. The results indicated that Hammett's acidity affected the yield of FAME. The following results were obtained: [TMEDAPS][HSO_4] (H_0 = 1.93, yielding 95%), [TMPDAPS][HSO_4] (H_0 = 1.90, yielding 95%), [TMHDAPS][HSO_4] (H_0 = 1.84, yielding 96%), and [MIMPS][HSO_4] (H_0 = 2.29, yielding 87%).

Figure S1 illustrates various families of ILs featuring different anions, such as NO_3, HSO_4, CF_3SO_3, CH_3SO_3, CF_3SO_3, and L-arginine. As shown, there is no clear correlation between Hammett's acidity and FAME yield. Therefore, the analysis indicated that Hammett acidity is not the sole parameter that should be considered to explain specific behavior.

There are additional physicochemical and structural properties that could enhance our understanding of the catalytic capacity of ILs. The literature indicates that Hammett's acidity is not a property often reported by researchers, since it is measured through spectrophotometric methods. This limitation hindered the analysis and correlation between the yield and the acidity of different ILs. Therefore, it is proposed to utilize a computational tool to predict the acidity of ILs with demonstrated catalytic capacity, as noted in the literature.

3.2. Predictive Acidity Analysis Using COSMO-RS

After analyzing the influence of Hammett acidity on the yield of fatty acid methyl ester synthesis, the acidity (α) of the hydrogen bonding in the interaction energies of ILs was determined using a predictive computational tool. The experimental acidity is well represented by a three-parameter model based on hydrogen-bonding energies (E_{HB}), electrostatic-misfit interactions (E_{misfit}), and van der Waals forces (E_{vdw}) in relation to the interaction energies of ILs (R^2 = 0.9441) as compared to the data of Kurnia et al.[27] Consequently, this model was employed to predict the acidity of the hydrogen bonds in ILs.

As illustrated in Figures S2 and S3 (Supporting Information), the relationship between the acidity calculated by COSMO-RS and the yield in FAMEs cannot be determined. Therefore, it is suggested that alternative criteria. Despite using experimental data from the literature along with the predictive model, sufficient information could not be gathered to define a relationship between the acidity and FAME yield.

3.2.1. Linear Model. Figure 5 shows data on the molecular interaction energies of ILs obtained from the COSMOtherm calculations. Initially, it was evident that this model type was inadequate for illustrating a correlation between the independent and dependent variables. This inadequacy arises because the interaction energies are derived from the average charge density distribution on each molecule's surface, making it impossible to obtain complete and comprehensive information about each IL. Therefore, it is suggested to implement more robust ML models that enhance the resolution between the independent and dependent variables. Proposed ML models include K-Nearest Neighbor (KNN), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), Decision Tree Regressor (DTR), and the Multilayer Perceptron (MLP) neural network model.

3.2.1.1. Correlation between Interaction Energies and Yield. The Pearson correlation matrix is presented in Table 1. Each cell in the matrix indicates the correlation coefficient between each pair of variables, with values ranging from −1 (perfect negative correlation) to 1 (perfect positive correlation). This visualization assists in identifying the variables with the strongest relationships. The correlations are classified as follows: high >0.7; moderate between 0.5 and 0.7; low between 0.3 and 0.5; and no correlation between 0 and 0.3.

The E_{misfit} energies of the cation and anion, although not highly correlated, significantly influenced the yield. Conversely, the E_{HB} and E_{misfit} of the cation and anion exhibited a strong correlation with each other. This is also evident in the heatmap (Figure 6), where the E_{misfit} energies displayed a greater color intensity, indicating the high correlation between these energies (shown in red).

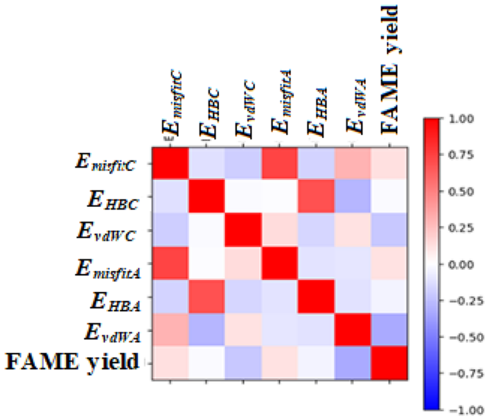


Figure 6. Correlation shown in heatmap.

3.2.1.2. Variance Inflation Factor. Multicollinearity refers to a situation in which one or more explanatory variables (predictors) in a multiple regression model are related to each other and, similarly, related to the response variable.[44] The variance inflation factor method was applied to address multicollinearity, which occurs when two independent variables are correlated with each other, complicating the interpretation of the results. In this study, no characteristics were removed because the *VIF* was not greater than 5 among

Table 2. Models Trained with the Cation and Anion Interaction Energies Approach for Ionic Liquids

	training data			test data		
models	MAE	RMSE	R^2	MAE	RMSE	R^2
LM	14.9240	394.7012	0.1827	16.5399	337.5208	0.2169
KNN	2.1600	60.7400	0.8742	12.2075	224.5055	0.4791
RFR	2.3611	60.8860	0.8739	12.5364	213.9963	0.5035
MLP	2.1617	60.7400	0.8742	11.1822	210.4090	0.5118
GBR	2.8277	61.5220	0.8726	10.7017	166.5553	0.6136
DTR	4.7534	87.3573	0.8191	10.9854	173.5125	0.5974

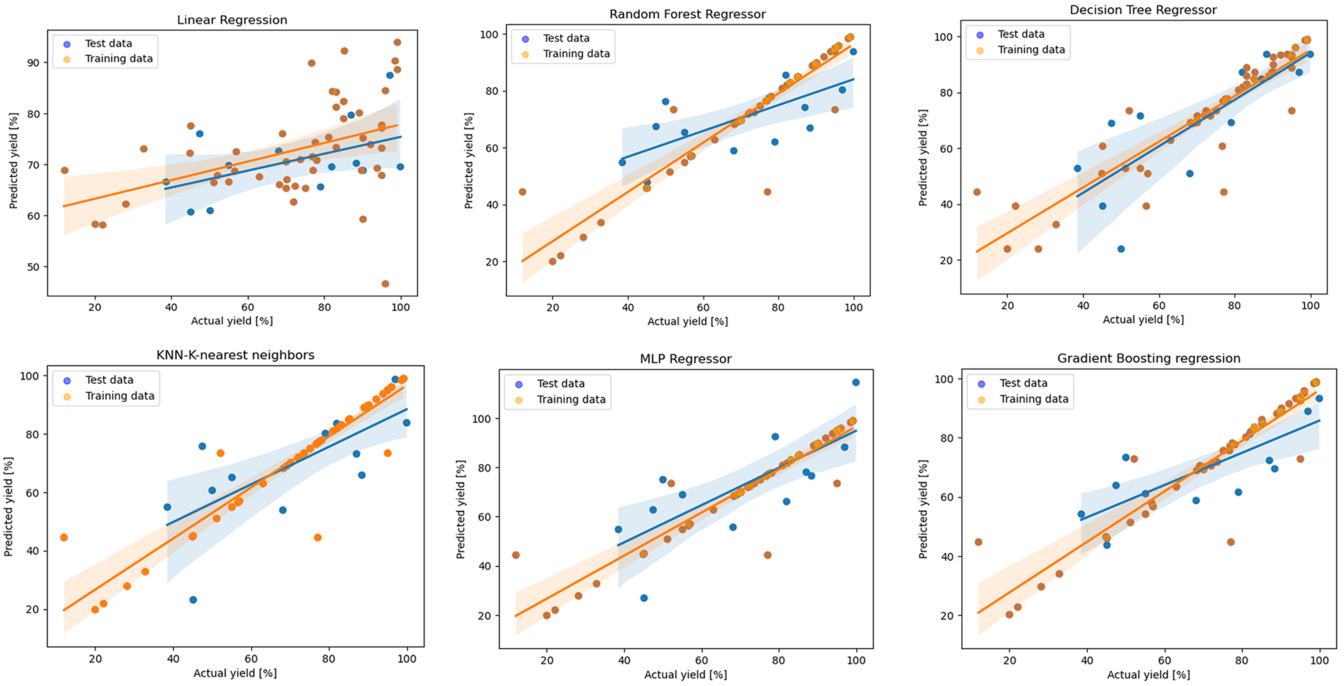


Figure 7. Predicting FAME performance using machine learning models.

two or more variables, as shown in Table S6 (Supporting Information).

3.3. Machine Learning Models. Table 2 presents the calculated values of the statistical parameters (R^2, RMSE and MAE) for the LM, KNN, RFR, MLP, GBR, and DTR models. As shown, the regression coefficients for the RFR, DTR, and GBR models were 0.5035, 0.5970, and 0.6136 for the test values, respectively.

The R^2 of 0.6136 indicates that 61.36% of the variation in the output variable can be explained by the input variables. The models that achieved the highest statistical resolution during the training method were KNN, MLP, and RFR; however, when applied to the test data, the performance was superior for the DTR and GBR models. Consequently, the GBR model was selected as the predictor model for the next stage as it demonstrated the best statistical performance during the test phase, exceptional performance during the training phase, and the lowest mean absolute error. This is illustrated in Figure 7, which shows that the GBR model for the test method had the best predictive performance based on the R^2, MAE, and RMSE statistical factors.

3.4. Sigma Profile

The σ profile diagram is divided into three regions. The $\sigma < -0.0082$ e/Å^2 region indicates that the substance has a strong ability to form hydrogen bonds and serves as a hydrogen bond donor region. The -0.0082 e/Å^2 $< \sigma <$ $+0.0082$ e/Å^2 region signifies molecular symmetry and is classified as an apolar region. In the $\sigma > +0.0082$ e/Å^2 region, this indicates that the substance has a strong capacity to accept hydrogen bonds and functions as a hydrogen bond acceptor region. The further the peak of the σ profile curve is from the line $\sigma = \pm 0.0082$ e/Å^2, and the larger the peak area, the stronger the corresponding property.[45]

To explain the interpretation of the sigma profile, the L-arginine cation molecule serves as a model (Figure 8). In these calculations, the COSMO continuous solvation model is employed to simulate a virtual conductive environment for the molecule, inducing an σ polarization charge density at the interface between the molecule and the conductor, specifically on the molecular surface. This results in a more polarized electron density compared to that in a vacuum. The 3D distribution of the polarization charges σ of each molecule is transformed into a surface composition function (σ-profile) obtained by using the COSMOtherm program. This σ-profile indicates the relative amount of σ-polarized surface on the molecule, giving detailed insights into the distribution of molecular polarity.[18] The σ profile of the L-arginine cation molecule features a peak in the strongly negative polar region (hydrogen bond donor) at -0.021 [e/Å^2], primarily associated with the guanidine N = H group, along with two peaks in the apolar region at -0.007 [e/Å^2] and $+0.003$ [e/Å^2], and a peak

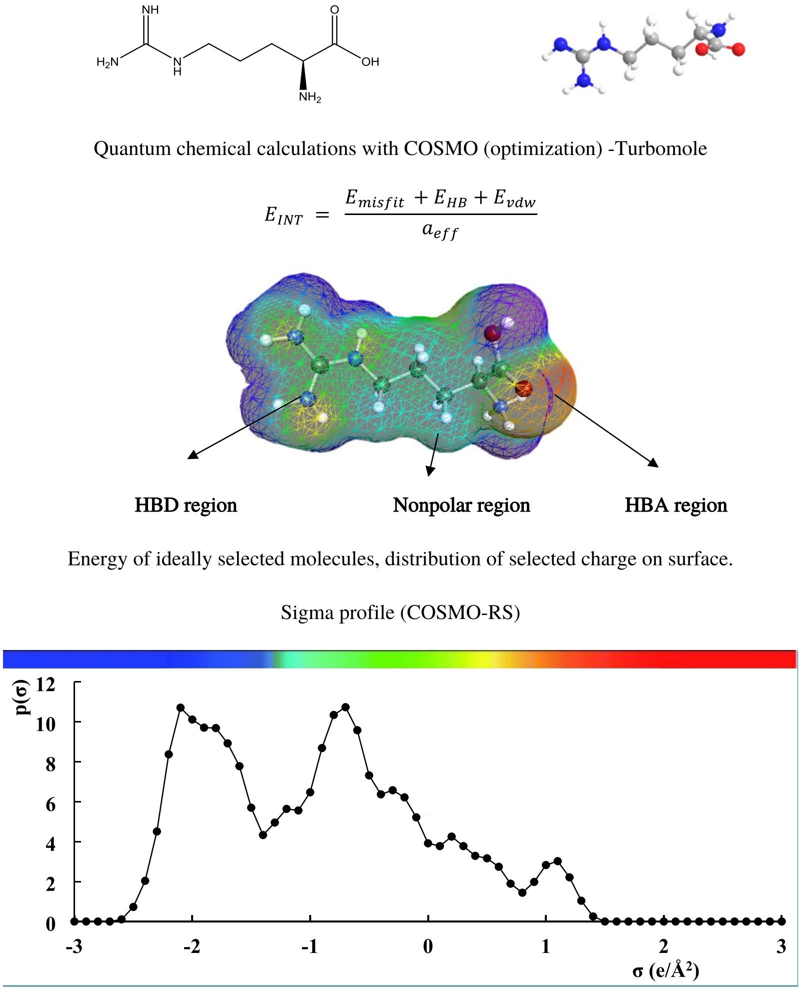

NH
O
H2N
N
H
OH
NH2
Quantum chemical calculations with COSMO (optimization) -Turbomole
$E_{INT} = \frac{E_{misfit} + E_{HB} + E_{vdw}}{a_{eff}}$
HBD region
Nonpolar region
HBA region
Energy of ideally selected molecules, distribution of selected charge on surface.
Sigma profile (COSMO-RS)
p(σ)
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
σ (e/Å²)

Figure 8. Relation between charge distribution and sigma profile for L-arginine molecule chemical structure.

in the strongly positively polar region (hydrogen bond acceptor) at +0.001 [e/Å^2], corresponding to the COOH carboxylic group. Additionally, it is evident that the molecule is more inclined to donate hydrogen bonds than to accept them.

3.4.1. Analysis of Sigma Profile Area Segments.

Once the interaction energies with the ML models were analyzed, the GBR model was applied in two different approaches (S_{CA} and $S_{C\text{-}A}$) that could impact the resolution of the FAME predictions, as developed by Torrecilla et al.[35] and Alkhatib et al.[36] using ML models for a similar purpose. At this stage, the model considered 75% of the data for training and 25% for testing. The parameters used in the model are *n_estimators = 100, max_depth = 4, max_features = log2, min_samples_leaf = 1, min_samples_split = 2, criterion = "squared_error", "squared_error"=0.1, tol = 0.0001.*

The GBR model, which analyzes the area under the σ profile curve of ILs, proved to be more efficient than the model that focuses on the interaction energies of ILs. Each of these energies independently represents an average of the energy contributions from the molecular interactions described by the surface charge density of the molecule. For instance, the hydrogen bond energy aggregates the energy densities in the hydrogen bond donor (E_{HBD}) region and the hydrogen bond acceptor (E_{HBA}) region. Thus, in analyzing the intermolecular interactions described by the surface energy densities in the three-dimensional distribution, it is assumed that these were insufficient, from a physical standpoint, to establish a correlation with the dependent variable (FAME yield).

Representing this information about the three-dimensional distribution of the molecules, the σ profile is derived in two dimensions. By utilization of the σ profile, it became possible to segment the distribution of energies into areas corresponding to energy types that are specific to the polar or apolar structures of the molecule. Five areas for each cation and anion were defined and included in the model independently, as illustrated in Figure 8. This set of energy areas under the σ profile curve represents the molecular interactions of each molecule and may provide a more comprehensive description of hydrogen-bonding interactions, as the molecular surface segments are arranged with distinct charge density in each area segment.[37,46]

In the calculated areas for different molecules, a fraction of the area derived information from E_{HBD}, $E_{HBD} + E_{misfit}$, $E_{misfit} + E_{vdW}$, $E_{misfit} + E_{HBA}$, and E_{HBA}. In other words, by employing this area approach, the ML models demonstrated greater accuracy than those relying solely on the information about molecular interactions (E_{HB}, E_{misfit}, and E_{vdW}), likely because this set of energy areas under the σ profile curve illustrated the molecular interactions of each molecule in a more comprehensive and realistic manner. Moreover, this analysis of area segments positively contributed to the ML model, as

Table 3. Models Trained with GBR Using Two Approaches

approaches	training data			test data		
models	MAE	RMSE	R^2	MAE	RMSE	R^2
(1) S_{CA}	2.0254	47.8152	0.9038	7.2301	81.0117	0.7955
(2) $S_{C\text{-}A}$	2.1948	48.0126	0.9034	5.8854	64.5684	0.8370

the GBR model performed better and was therefore more accurate according to the defined statistical criteria.

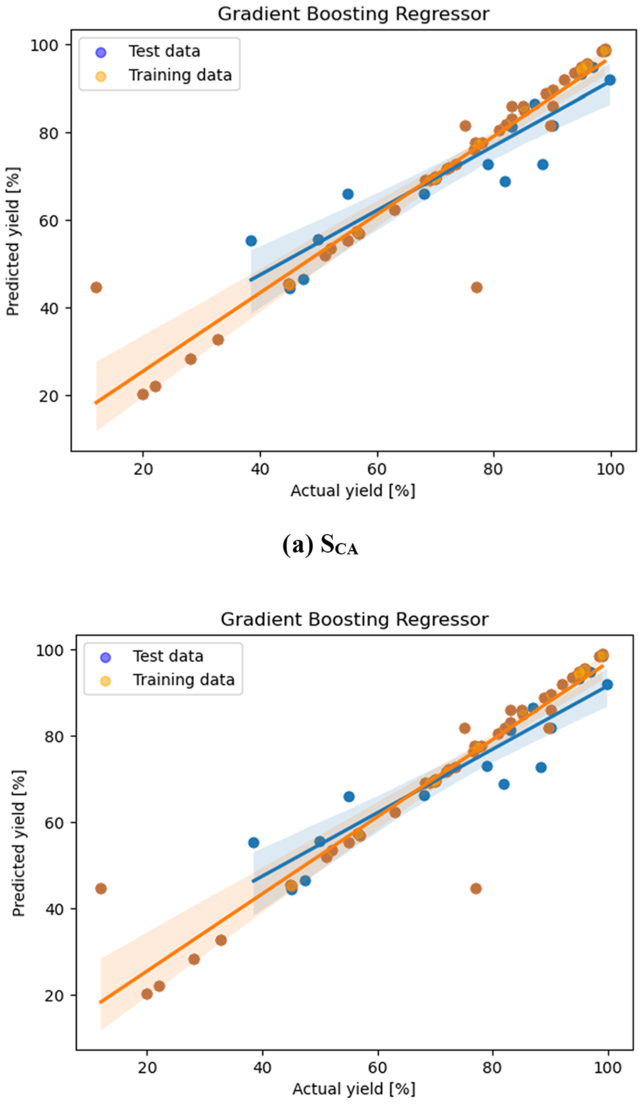
The S_{CA} and $S_{C\text{-}A}$ approaches are also discussed. Although these methods are statistically different, the $S_{C\text{-}A}$ approach has a greater impact on predicting FAME. This indicates that the molecular interaction energies obtained from each area segment calculated in the S_{CA} approach represent each polar or apolar region of the molecule in a single area formed by the respective cation and anion, resulting in five areas that were used as inputs to train and validate the model. However, in the $S_{C\text{-}A}$ approach, the energy area segments possess their own areas and energy charge density. Therefore, this approach allowed for the analysis of 10 area segments as the input for the model.

By analyzing the results and using the statistical parameters MAE, RMSE, and R^2, it is evident that the GBR model (3) achieves a 22.34% improvement in prediction during the testing phase compared to the GBR model (1), as shown in Table 3. Additionally, this demonstrates better statistical performance during the training phase. Conversely, when examining the GBR model alongside the S_{CA} and $S_{C\text{-}A}$ approaches, as illustrated in Figure 9a,b, respectively, it is observed that the GBR model (2) provides a 4.15% higher prediction, indicated by the R^2, compared to the GBR model (1). This is further demonstrated in Table 3, which shows that the $S_{C\text{-}A}$ approach offers a better fit for two data sets during the testing phase.

The ILs and their corresponding predictions in FAME are provided in Table S7 of the Supporting Information. ILs based on amino acids, inorganic acids (HSO_4, HNO_3), and carboxylic acids such as acetic and propanoic acids exhibit high performance predictions using the ML model. Following the criteria outlined in the methodology, the ILs [L-arginine][Acetate], [L-arginine][HSO_3], and [L-arginine]-[NO_3] were chosen for their synthesis and characterization. The characterization of ionic liquids can be found in the Supporting Information (Figures S4−S6).

4. CONCLUSION

In this study, a machine learning model was developed to predict fatty acid methyl esters in a series of ionic liquids utilizing molecular data from COSMO calculations. The results presented highlight the effectiveness of two machine learning models in predicting the performance of fatty acid methyl esters (FAME) in ionic liquids. The Gradient Boosting Regressor model using the $S_{C\text{-}A}$ approach demonstrated superior performance and statistical accuracy in FAME prediction. Both the COSMO-RS thermodynamic model and the machine learning models can be combined to forecast the performance of fatty acid methyl esters (FAME). The ionic liquids [L-arginine][Acetate], [L-arginine][HSO_3], and [L-arginine][NO_3] were synthesized and characterized using nuclear magnetic resonance (NMR ^{1}H, ^{13}C), along with the prediction criteria from the Gradient Boosting Regressor machine learning model on the Jupyter Notebook computing platform using the Python programming language. The optimized GBR ($S_{C\text{-}A}$) model can reliably and accurately predict ionic liquid FAMEs and assist in selecting ionic liquids with a suitable catalytic capacity for FAME synthesis.



(a) S_{CA}

(b) $S_{C\text{-}A}$

Figure 9. GBR model using the (a) S_{CA} and (b) $S_{C\text{-}A}$ approach.

ASSOCIATED CONTENT

Data Availability Statement

Data will be made available on request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsengineeringau.5c00098.

Study on protic ionic liquids used as catalysts in transesterification; tables with data on predictive acidity, interaction energies, and areas under the sigma profile curves of these ionic liquids; NMR spectra for liquids such as [L-arginine][Acetate], [L-arginine][NO_3], and [L-arginine][HSO_3], detailing the chemical shifts; graphs and predictions of FAME (fatty acid methyl esters) yield generated by machine learning models, analyzing the acidity of anions and cations (PDF)

AUTHOR INFORMATION

Corresponding Author

Pedro Felipe Arce − *Chemical Engineering Department, Engineering School of Lorena, University of São Paulo (USP), Lorena, São Paulo 12602-810, Brazil;* orcid.org/0000-0002-4687-5297; Phone: 55-12-31595326; Email: parce@usp.br

Authors

Luis Alberto Gallo-García − *Chemical Engineering Department, Engineering School of Lorena, University of São Paulo (USP), Lorena, São Paulo 12602-810, Brazil; CICECO—Aveiro Institute of Materials, Department of Chemistry and CESAM—Centre for Environmental and Marine Studies, Department of Environment and Planning, University of Aveiro, Aveiro 3810-193, Portugal*

Pedro J. Carvalho − *CICECO—Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Aveiro 3810-193, Portugal;* orcid.org/0000-0002-1943-0006

Maria Isabel da Silva Nunes − *CESAM—Centre for Environmental and Marine Studies, Department of Environment and Planning, University of Aveiro, Aveiro 3810-193, Portugal*

Nian Vieira Freire − *Chemical Engineering Department, Engineering School of Lorena, University of São Paulo (USP), Lorena, São Paulo 12602-810, Brazil;* orcid.org/0000-0003-1257-4928

Alessandro Cazonatto Galvao − *Laboratory ApTher—Applied Thermophysics, Department of Food and Chemical Engineering, Santa Catarina State University (UDESC), Pinhalzinho, Santa Catarina 89870-000, Brazil;* orcid.org/0000-0002-8255-4511

Daniela Helena Pelegrine Guimarães − *Chemical Engineering Department, Engineering School of Lorena, University of São Paulo (USP), Lorena, São Paulo 12602-810, Brazil;* orcid.org/0000-0002-4797-1168

Complete contact information is available at: https://pubs.acs.org/10.1021/acsengineeringau.5c00098

Author Contributions

CRediT: **Luis Alberto Gallo Garcı́a** data curation, formal analysis, methodology, software, validation, writing - original draft; **Pedro J Carvalho** and **Maria Isabel da Silva Nunes** investigation, project administration, resources, supervision, writing - review & editing; **Nian Vieira Freire** data curation, formal analysis, software, validation, writing - original draft; **Alessandro Cazonatto Galvão** project administration, supervision, writing - review & editing; **Daniela Helena Pelegrine Guimarães** investigation, project administration, resources, supervision, writing - review & editing; **Pedro Felipe Arce** investigation, project administration, resources, supervision, writing - review & editing.

Funding

The Article Processing Charge for the publication of this research was funded by the Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES), Brazil (ROR identifier: 00x0ma614).

Notes

I would like to declare on behalf of my coauthors that the work described was original research that has not been published previously, and not under consideration for publication elsewhere, in whole or in part.

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

L.A.G.-G. thanks the Ministerio de Ciencia Tecnología e Innovación (MinCiencias) of Colombia, Call No. 885 of 2020, for the doctoral scholarship. P.J.C. and M.I.d.S.N. thank the FCT/MCTES for the financial support to CICECO—Aveiro Institute of Materials (UIDB/50011/2020, UIDP/50011/2020 and LA/P/0006/2020) and to CESAM (UIDP/50017/2020 + UIDB/50017/2020 + LA/P/0094/2020). N.V.F. thanks the Financial supports of CAPES (Brazil), Finance Code 001. P.F.A. and D.H.P.G. thank the financial support of the Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP (Brazil) (FAPESP processes: 2020/15809-3, 2022/11043-1, and 2024/14835-1. The Article Processing Fee for the publication of this research was funded by the Coordination for the Improvement of Higher Education Personnel - CAPES (identifier ROR: 00x0ma614).)

REFERENCES

(1) Guo, M.; Jiang, W.; Chen, C.; Qu, S.; Lu, J.; Yi, W.; Ding, J. Process optimization of biodiesel production from waste cooking oil by esterification of free fatty acids using La3+/ZnO-TiO2 photocatalyst. *Energy Convers. Manag.* **2021**, *229*, No. 113745.

(2) Islam, A.; Teo, S. H.; Islam, M. T.; Mondal, A. H.; Mahmud, H.; Ahmed, S.; Ibrahim, M.; Taufiq-Yap, Y. H.; Hossain, M. L.; Sheikh, M. C.; et al. Harnessing visible light for sustainable biodiesel production with Ni/Si/MgO photocatalyst. *Renewable Sustainable Energy* **2025**, *208*, No. 115033.

(3) Yusuf, B. O.; Oladepo, S. A.; Ganiyu, S. A. Efficient and sustainable Biodiesel Production via Transesterification: catalysts and operating conditions. *Catalysts* **2024**, *14*, No. 581.

(4) Mandari, V.; Devarai, S. K. Biodiesel production using homogeneous, heterogeneous, and enzyme catalysts via transesterification and esterification reactions: A critical review. *BioEnergy Res.* **2022**, *15*, 935−961.

(5) Endalew, A. K.; Kiros, Y.; Zanzi, R. Inorganic heterogeneous catalysts for biodiesel production from vegetable oils. *Biomass Bioenerg.* **2011**, *35*, 3787−3809.

(6) Ngomade, S. B. L.; Fotsop, C. G.; Bhonsle, A. K.; Rawat, N.; Gupta, P.; Singh, R.; Tchummegne, I. K.; Singh, R. K.; Atray, N. Pilot-scale optimization of enhanced biodiesel production from high FFA Podocarpus falcatus oil via simultaneous esterification and transesterification assisted by zirconia-supported ZSM-5. *Chem. Eng. Res. Des.* **2024**, *209*, 52−56.

(7) Di Serio, M.; Tesser, R.; Pengmei, L.; Santacesaria, E. Heterogeneous catalysts for biodiesel production. *Energy Fuels* **2008**, *22*, 207−217.

(8) Ding, H.; Ye, W.; Wang, Y.; Wang, X.; Li, L.; Liu, D.; Gui, J.; Song, C.; Ji, N. Process intensification of transesterification for biodiesel production from palm oil: Microwave irradiation on transesterification reaction catalyzed by acidic imidazolium ionic liquids. *Energy* **2018**, *144*, 957−967.

(9) Ong, H. C.; Tiong, Y. W.; Goh, B. H. H.; Gan, Y. Y.; Mofijur, M.; Fattah, I. M. R.; Chong, C. T.; Alam, M. A.; Lee, H. V.; Silitonga,

A. S.; Mahlia, T. Recent advances in biodiesel production from agricultural products and microalgae using ionic liquids: Opportunities and challenges. *Energy Convers. Manag.* **2021**, *228*, No. 113647.

(10) Huang, Z.; Yang, Y. C.; Wang, X.; Cai, R.; Han, B. Biodiesel synthesis through soybean oil transesterification using choline-based amino acid ionic liquids as catalysts. *Ind. Crop. Prod.* **2023**, *208*, No. 117869.

(11) Li, J.; Guo, Z. Structure evolution of synthetic amino acids-derived basic ionic liquids for catalytic production of biodiesel. *ACS Sustainable Chem. Eng.* **2017**, *5*, 1237−1247.

(12) Fan, M.; Huang, J.; Yang, J.; Zhang, P. Biodiesel production by transesterification catalyzed by an efficient choline ionic liquid catalyst. *Appl. Energy* **2013**, *108*, 333−339.

(13) Han, M.; Yi, W.; Wu, Q.; Liu, Y.; Hong, Y.; Wang, D. Preparation of biodiesel from waste oils catalyzed by a Brønsted acidic ionic liquid. *Biores. Technol.* **2009**, *100*, 2308−2310.

(14) O'Connor, S.; Pillai, S. C.; Ehimen, E.; Bartlett, J. Production of Biodiesel Using Ionic Liquids. In *Nanotechnology-Based Industrial Applications of Ionic Liquids*; Inamuddin; Asiri, A., Eds.; Springer: Cham, NY, 2020 DOI: 10.1007/978-3-030-44995-7_12.

(15) Zhang, Y.; Sun, S. A review on biodiesel production using basic ionic liquids as catalysts. *Ind. Crops. Prod.* **2023**, *202*, No. 117099.

(16) Eckert, F.; Klamt, A. Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J.* **2002**, *48*, 369−385.

(17) Klamt, A.; Eckert, F. COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. *Fluid Phase Equilib.* **2000**, *172*, 43−72.

(18) Palomar, J.; Torrecilla, J. S.; Lemus, J.; Ferro, V. R.; Rodríguez, F. Prediction of non-ideal behavior of polarity/polarizability scales of solvent mixtures by integration of a novel COSMO-RS molecular descriptor and neural networks. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5967−5975.

(19) Palomar, J.; Torrecilla, J. S.; Lemus, J.; Ferro, V. R.; Rodríguez, F. A. A COSMO-RS based guide to analyze/quantify the polarity of ionic liquids and their mixtures with organic cosolvents. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1991−2000.

(20) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255−260.

(21) Kim, S.; Seo, J.; Kim, S. Machine learning technologies in the supply chain management research of biodiesel: a review. *Energies* **2024**, *17*, 1316.

(22) Diedenhofen, M.; Klamt, A. COSMO-RS as a tool for property prediction of IL mixtures—A review. *Fluid Phase Equilib.* **2010**, *294*, 31−38.

(23) Klamt, A. The COSMO and COSMO-RS solvation models. *Wiley Interdiscip. Rev.:Comput. Mol. Sci.* **2011**, *1*, 699−709.

(24) Hajipour, A. R.; Rafiee, F. Acidic bronsted ionic liquids. *Org. Prep. Proced. Int.* **2010**, *42*, 285−362.

(25) Amarasekara, A. S. Acidic ionic liquids. *Chem. Rev.* **2016**, *116*, 6133−6183.

(26) Gao, j.; Zhu, y.; Liu, w.; Jiang, S.; Zhang, J.; Ma, W. Hydrogen bonds in disulfonic-functionalized acid ionic liquids for efficient biodiesel synthesis. *ACS Omega* **2020**, *5*, 12110−12118.

(27) Kurnia, K. A.; Lima, F.; Cláudio, A. F. M.; Coutinho, J. A. P.; Freire, M. G. Hydrogen-bond acidity of ionic liquids: an extended scale. *Phys. Chem. Chem. Phys.* **2015**, *17*, 18980−18990.

(28) Shrestha, N. Detecting multicollinearity in regression analysis. *Am. J. Appl. Math. Stat.* **2020**, *8*, 39−42.

(29) Bascuñana, J.; León, S.; González-Miquel, M.; González, E. J.; Ramírez, J. Impact of Jupyter Notebook as a tool to enhance the learning process in chemical engineering modules. *Educ. Chem. Eng.* **2023**, *44*, 155−163.

(30) Otchere, D. A.; Ganat, T. O. A.; Ojero, J. O.; Tackie-Otoo, B. N.; Taki, M. Y. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *J. Petrol. Sci. Eng.* **2022**, *208*, No. 109244.

(31) Sumayli, A. Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models. *Arab. J. Chem.* **2023**, *16*, No. 104833.

(32) Tyralis, H.; Papacharalampous, G. Boosting algorithms in energy research: a systematic review. *Neural Comput. Appl.* **2021**, *33*, 14101−14117.

(33) Abranches, D. O.; Maginn, E. J.; Colón, Y. J. Boosting graph neural networks with molecular mechanics: A case study of sigma profile prediction. *J. Chem. Theory Comput.* **2023**, *19*, 9318−9328.

(34) Mullins, E.; Liu, Y. A.; Ghaderi, A.; Fast, S. R. Sigma profile database for predicting solid solubility in pure and mixed solvent mixtures for organic pharmacological compounds with COSMO-based thermodynamic methods. *Ind. Eng. Chem. Res.* **2008**, *47*, 1707−1725.

(35) Torrecilla, J. S.; Palomar, J.; Lemus, J.; Rodríguez, F. A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids. *Green Chem.* **2010**, *12*, 123−134.

(36) Alkhatib, I. I. I.; Albà, C. G.; Darwish, A. S.; Llovell, F.; Vega, L. F. Searching for sustainable refrigerants by bridging molecular modeling with machine learning. *Ind. Eng. Chem. Res.* **2022**, *61*, 7414−7429.

(37) Hsieh, C. M.; Sandler, S. I.; Lin, S. T. Improvements of COSMO-SAC for vapor−liquid and liquid−liquid equilibrium predictions. *Fluid Phase Equilib.* **2010**, *297*, 90−97.

(38) Chicco, D.; Warrens, M. J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, No. e623.

(39) Sharma, G.; Singh, D.; Rajamani, S.; Gardas, R. L. Influence of Alkyl Substituent on Optical Properties of Carboxylate-Based Protic Ionic Liquids. *ChemistrySelect* **2017**, *2*, 10091−10096.

(40) Martins, M. A. R.; Sharma, G.; Pinho, S. P.; Gardas, R. L.; Coutinho, J. A. P.; Carvalho, P. J. Selection and characterization of non-ideal ionic liquids mixtures to be used in CO2 capture. *Fluid Phase Equilib.* **2020**, *518*, No. 112621.

(41) Masri, A. N.; Mutalib, M. I. A.; Aminuddin, N. F.; Lévêque, J. Novel SO3H-functionalized dicationic ionic liquids − A comparative study for esterification reaction by ultrasound cavitation and mechanical stirring for biodiesel production. *Sep. Purif. Technol.* **2018**, *196*, 106−114.

(42) Tankov, I.; Mustafa, Z.; Nikolova, R.; Veli, A.; Yankova, R. Biodiesel (methyl oleate) synthesis in the presence of pyridinium and aminotriazolium acidic ionic liquids: Kinetic, thermodynamic studies. *Fuel* **2022**, *307*, No. 121876.

(43) Li, M.; Chen, J.; Li, L.; Ye, C.; Lin, X.; Qiu, T. Novel multi−SO3H functionalized ionic liquids as highly efficient catalyst for synthesis of biodiesel. *Green Energy Environ.* **2021**, *6*, 271−282.

(44) Akinwande, M. O.; Dikko, H. G.; Samson, A. Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open J. Stat.* **2015**, *05*, 754−767.

(45) Yin, Y.; Tang, Z.; Liu, W.; Li, X.; Sun, C.; Zhang, W.; Xu, X. Study on Vapor−Liquid Equilibrium of L-Lysine, L-Arginine, and L-Threonine Aqueous Solutions. *J. Chem. Eng. Data* **2024**, *69*, 2783−2792.

(46) Mahmoudabadi, S. Z.; Pazuki, G. Investigation of COSMO-SAC model for solubility and cocrystal formation of pharmaceutical compounds. *Sci. Rep.* **2020**, *10*, No. 19879.