Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Predicting Surface Tension of Organic Molecules using COSMO-RS Theory and Machine Learning

Flora Esposito[a], Ulderico Di Caprio[a], Bruno Rodrigues[c], Florence H. Vermeire[b], Idelfonso B.R. Nogueira[c], and M.Enis Leblebici[a,*]

[a] Center for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, 3590 Diepenbeek, Belgium
[b] KU Leuven, Department of Chemical Engineering, Celestijnenlaan 200F-bus 2424, Leuven 3001, Belgium
[c] Department of Chemical Engineering; Norwegian University of Science and Technology, Gløshaugen, Trondheim, 7034, Norway
* Corresponding Author: muminenisleblebici@kuleuven.be.

ABSTRACT

Surface tension is a fundamental property at the liquid/gas interface, influencing phenomena such as capillary action, droplet formation, and interfacial behavior in chemical engineering processes. Despite its significance, experimental determination of surface tension is time-intensive and impractical for in silico-designed compounds. Predictive models are essential for bridging this gap. This study expands on Gaudin's COSMO-RS-based model, which assumes uniform molecular orientation at the surface, by testing its predictive capability across broader temperatures (5–50°C) and developing a hybrid model combining first-principle and machine learning insights to improve Gaudin's model predictions. The HM employs a serial configuration where COSMO-RS predictions serve as inputs alongside molecular descriptors, derived using the Mordred library. SHAP analysis guides feature selection, enhancing model interpretability. An artificial neural network refines predictions, optimized via Bayesian optimization for architecture and hyperparameters. Using a diverse dataset of 85 organic molecules, the HM demonstrates improved accuracy, reducing various metrics compared to standalone first-principle predictions.

Keywords: COSMO-RS, Machine Learning, Hybrid Modeling, Surface tension, First-Principle modeling.

INTRODUCTION

Surface tension is a key property at the liquid/gas interface and plays an important role in various chemical engineering processes, including liquid flow and transport through porous media. This property is responsible for various phenomena, such as capillary action, droplet formation, and the behavior of liquid interfaces. Surface tension has practical applications in various fields, e.g., thin layer technologies, sol–gel technologies, phase separation techniques, drug carriers and treatment of raw materials [1]. Experimental measurement results for this property require specialized equipment and are time-consuming, therefore predictive models for surface tension are essential. Moreover, significantly different surface tension values for a given compound can arise from the method used (e.g., Wilhelmy plate, Du Noüy ring, maximum bubble pressure, dropweight, or hanging-drop) and the liquid purity [2]. Additionally, a transition to a greener chemical industry requires the development of innovative and sustainable solvents. Utilizing an in-silico model enables the pre-screening of these materials without the need for experimental procedures, thereby accelerating process development. Therefore, reliably predicting surface tension values for candidate compounds is essential. Numerous predictive models have been proposed in the literature for surface tension, e.g. neural network regression and graph machine regression [2], Quantitative Structure-Property Relationship (QSPR) models [3], combination of group-contribution and machine learning (ML) models [4]. These rely on large datasets and maintain reliability only within the boundaries of their training data. Moreover, they do not incorporate any underlying physical principles of the system. Quantum chemistry-based models are highly effective in predicting a wide range of physicochemical properties of organic compounds. The Conductor-like Screening Model for Real Solvents (COSMO-RS) method,

introduced by Andreas Klamt in 1995, integrates unimolecular quantum chemical calculations with exact statistical thermodynamics [5]. This approach enables the prediction of chemical potentials in liquids, which can be used to derive various thermodynamic equilibrium properties, without relying on experimental data. For example, COSMO-RS can be used to compute interfacial tension, which represents the free energy per unit surface area required to maintain an interface between two condensed phases. However, accurately predicting surface tension remains a significant challenge for COSMO-RS [6]. Recently, Gaudin proposed a COSMO-RS based model to predict surface tension (σ) of pure liquids [7]. Gaudin's approach is employed in this work and explained in the Methodology paragraph of this manuscript. Briefly, this method assumes uniform molecular orientation, meaning that every surface segment has an equal probability of being exposed. While originally tested on a limited set of molecules at 20°C, this study aims to (1) evaluate its predictive performance across a broader temperature range and (2) develop a Hybrid Model (HM) that combines COSMO-RS predictions with molecular descriptors to estimate experimental surface tension for more robust predictions. By integrating domain knowledge, HMs improve system understanding while requiring less data than purely ML models. This approach enhances reliability and offers a high benefit-to-cost ratio for complex systems [8]. In this work, a serial HM is employed, where the first-principle model (FPM) predictions are used as an input for the ML model together with molecular descriptors. The ML model then predicts the experimental surface tension value.

METHODOLOGY

First Principle predictions

Gaudin proposed a COSMO-RS based model to predict surface tension (σ_{sim}) of pure liquids [7]. The model is described by

$$\sigma_{sim} = -\frac{\Delta G_{sim}}{S_{sim}}, \quad (1)$$

where σ_{sim} (mN/m) is the simulated surface tension, ΔG_{sim} (kcal/mol) is the COSMO-RS simulated Gibbs free energy of self-solvation and S_{sim} (Å^2) is the molecular flat surface. The implicit assumption of this model is that the molecular orientation at the surface is neglected: any surface segment has the same probability to be exposed at the surface. However, this assumption may not be true for compounds where specific molecular orientations are energetically favored. In such cases, surface segments with lower desolvation energy are more likely to be exposed. Both the authors of this work and Gaudin suggest that deviations in surface tension predictions stem from this simplified hypothesis.

Data gathering

The experimental surface tension data, used for model training, has been retrieved online (surface-tension.de website). The dataset contains values for 85 organic molecules. These molecules vary from small, simple compounds to larger, complex structures, with molecular weights ranging from 18.02 g/mol (water) to 260.8 g/mol (hexachlorobutadiene). The dataset includes hydrocarbons, halogenated compounds, alcohols, and nitrogen-containing molecules. Additionally, there are compounds with oxygen-containing functional groups, such as ethers, esters and ketones. The molecules exhibit a broad range of polarities, with polar compounds contrasting nonpolar hydrocarbons. Structural complexity also varies, including linear alkanes, aromatic rings, and cyclic compounds. This chemical diversity makes the dataset suitable for exploring a wide range of physical and chemical properties. The Gibbs self-solvation free energy (ΔG_{sim}) and molecular surfaces (S_{sim}) were caulcated from COSMOTherm software (version 24.0.0) using the BP-TVZP parameterisation, as suggested by Gaudin. Simulations were conducted over a temperature range of 5 to 50 °C, with increments of 5 °C for ΔG_{sim} predictions. Molecular flat surface data were extracted from the COSMO file of each molecule. A comparison between the simulated surface tensions in this study and those from Gaudin's work revealed strong agreement. Error metrics have been calculated point by point for ΔG_{sim}, S_{sim} and σ_{sim}, at T = 20 C, at which Gaudin results are available. The maximum absolute error observed was 0.66 kcal/mol for ΔG_{sim}, 8.41 Å^2 for S_{sim} , and 6.74 mN/m for the σ_{sim}. The maximum absolute percentage error observed was 8% for ΔG_{sim}, 5% on the S_{sim} and 11% for σ_{sim}. The coefficient of determination (R^2) for σ_{sim}, S_{sim} and ΔG_{sim} is 0.989, 0.999 and 0.991 respectively. The small deviations observed could be due to slightly different parametrization in older versions of COSMOTherm or to the chosen molecules conformer.

Features selection and data preprocessing

The developed model is a HM where the FPM is the one reported in Eq 1 and the ML part is composed by an Artificial Neural Network (ANN). The starting point of the ML part is the Simplified Molecular Input Line Entry System (SMILES) of the molecules. SMILES strings were stored from COSMOTherm along with simulation results, but can be easily retrieved from RDKit library as well. The SMILES strings are processed using the RDKit library (version 2024.9.4) in Python to generate chemical structures, substructures, and conformations, represented as text-based connection tables (MOL files). The resulting MOL files are subsequently analyzed with the Mordred library (version 1.2.0) in Python to compute molecular descriptors [9]. 37 2D descriptors are calculated, among which four are selected as inputs for the model. The

calculated descriptors concern the molecule aromaticity, structure, geometry, polarity. They were chosen for their potential influence on surface tension. Deviations in flat surface measurements are already expected due to polarity and may also result from molecular geometry, weight, hydrophobicity, and aromaticity. Feature selection is conducted iteratively using SHAP analysis, a game-theoretic approach that explains machine learning model predictions by assigning contribution values (SHAP values) to each feature [10]. This method helps identify which features drive the predictions and whether their influence positively or negatively affects the predicted surface tension. The SHAP (version 0.46.0.) library in Python has been used for the analysis. The selection process begins with training a neural network with a randomly chosen structure (e.g., three hidden layers with 30 nodes per layer) and evaluating its performance on test set metrics. In each iteration, the least important feature identified through SHAP analysis is removed, and the model is retrained. This process continues until the model achieves satisfactory performance, i.e. a MSE lower than that of a standalone FPM. This procedure is shown in Figure 1. The final selected features are the atom and bond polarizability, the average molecular weight, and the number of double bonds in Kekulized structure. In addition to these features, the COSMO-RS simulated surface tensions and the corresponding temperature are also used as model input. The dataset is therefore composed of molecular descriptors, temperatures, and surface tension calculations at every temperature. Given 85 molecules and temperatures ranging from 5 to 50°C with a step of 5°C, the dataset is composed of 850 examples and 6 features: simulated surface tension, temperature, atom and bond polarizability, average molecular weight and the number of double bonds in kekulized structure. The molecules have been organised by CAS Number and the test set has been randomly selected, while making sure that the molecules appearing in the test set are not included in the training set. For simplicity, the first 680 rows, e.g. the first 68 molecules of the dataset are used as training set and the remaining 170 rows, e.g. 17 molecules are used as test set. The data is then standardized prior to model training using a standard scaler in scikit learn (version 1.6.1) library.

Model development

The workflow of the HM is given in Figure 2. Once the FPM calculations are completed, literature data gathered and features selected, the ANN can be trained for predicting experimental values of surface tensions from 5 to 50°C. The trained neural network is built and optimized in Keras library (version 3.5.0) in Python. The loss function is represented by the mean squared error (MSE). Activation function is Leaky Rectified Linear Unit (Leaky ReLU) with α = 0.01 for all layers, except the output one which has a linear activation function. The number of layers, neurons per layers and learning rate of the ANN are optimized through a Bayesian optimization (BO). The number of hidden layers to be chosen by the optimizer ranges from 2 to 5, while the number of nodes per layer ranges from 5 to 100 with a step of 3. The learning rate minimum and maximum boundaries were 10^{-4} and 10^{-2} respectively. During BO, the dataset is divided into training, validation, and test subsets. The test set comprises the last 17 molecules in the dataset, while the validation set is formed by randomly selecting 7 molecules (corresponding to 70 rows) from the training set across temperatures ranging from 5 to 50 °C. The loss function is the MSE on the validation set. After identifying the optimal network structure, the molecules initially used for validation are reintegrated into the training set, ensuring that the test set remains entirely unseen by both the model and the optimizer throughout the process. The proposed structure is formed by three hidden layers. The number of nodes per layer is 17,65,59 respectively and the learning rate is 0.00184. The model prediction accuracy is evaluated on various metrics both on training and test set e.g. the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), MSE, the Root Mean Squared Error (RMSE), and the R^2. The same evaluation metrics are also used to compare the FPM standalone prediction accuracy.

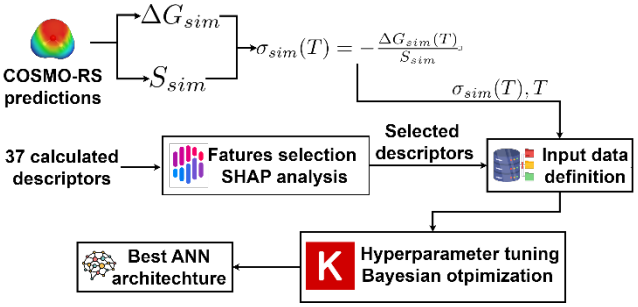

Figure 1: Training workflow. Self-solvation energies and molecular flat surfaces are obtained from COSMO-RS predictions, and surface tension is subsequently calculated. A total of 37 molecular descriptors are computed. Feature selection is carried out using SHAP analysis through an iterative process. Upon completion, the selected descriptors, simulated surface tension, and corresponding temperature constitute the input dataset. Hyperparameter tuning is then conducted to identify the optimal model architecture.

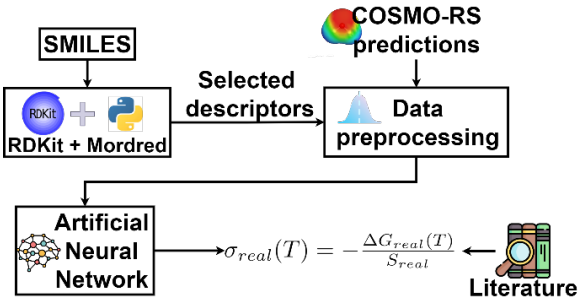

Figure 2: HM model workflow. Mol files are retrieved from SMILES in RDKit library. Molecular descriptors are calculated from Mordred library. Surface tensions, temperatures and descriptors are provided to an ANN which can predict temperature-dependent surface tensions experimental data.

RESULTS

Model prediction accuracy

Figure 3 shows a comparison of the predicted surface tension using Gaudin standalone model, the developed HM and experimental values. The left plot shows the parity plot for a range of temperatures from 5 to 50°C, while the right one shows predictions for a temperature of 25°C for the HM only, middle point of the selected temperatures. Looking at the standalone FPM, for some, less polar molecules, e.g. cyclohexane, 1-chlorobutane, dodecane, carbon tetrachloride, 1-nitropropane, Gaudin approximation seems to hold. Nevertheless, some molecules that show strong hydrogen-bonding association may orient in a specific way at the interface, with consequent deviations from experimental measurements. Overestimated surface tension is indeed evident for molecules above the 20% error line. These are methanol, ethanol, water, formamide, glycol, propanol and 2-propanol. 5 of these are also cited as overestimated outliers in Gaudin's paper. Overall, apart from the molecules mentioned above, the standalone FPM slightly underestimates surface tension. This trend is visible in the clustering of green data points below the parity line. Focusing on the HM model predictions, instead, all the points, for both training and test sets, fall within the 20% error region, showing the improved accuracy of the HM compared to the FPM. The plots demonstrate that the ANN effectively corrects the deviations of the FPM, even for the aforementioned polar molecules that exhibit significant discrepancies from the experimental values. Table 1 presents Gaudin's standalone model metrics for predicting surface tension, along with the percentage improvement in these metrics when the FPM is compared to the HM test set performance. The decrease of all the error metrics and the outstanding increase of the R^2 confirm that the HM offers a significant improvement in predictive performance.

Table 1: FPM metrics and corresponding percentage improvement achieved when compared to the HM test set metrics.

Metric	FPM	Improvement (%)
MAE	5.99 mN/m	-75.6%
MAPE	17.95 %	-78.9%
MSE	70.25 mN^2/m^2	-95.0%
RMSE	8.3816 mN/m	-77.8%
R^2	0.4281	+111.0%

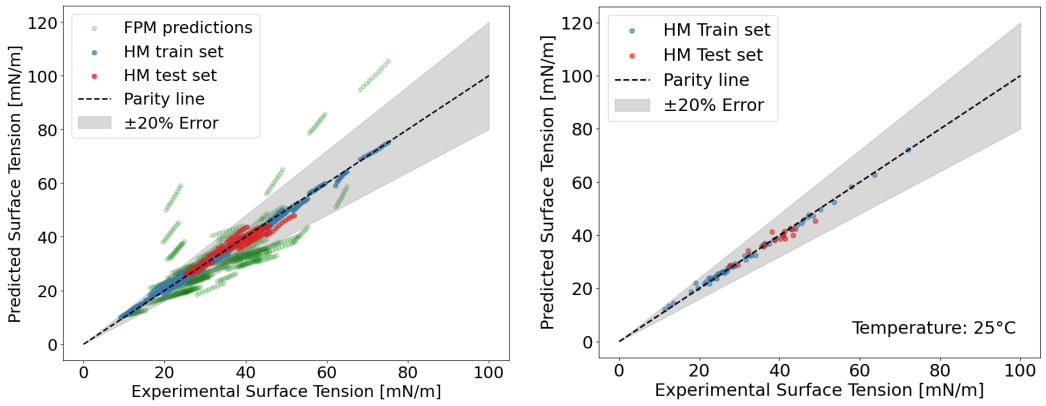

Figure 3: Comparison between the FPM, the HM developed in this work and experimental data. The left plot shows results of surface tension calculations from 5 to 50 °C. The green points represent the FPM model predictions. Blue and red points represent the HM train and test dataset, respectively. The black dashed line represents the ideal scenario where predicted values perfectly match experimental values, indicating a model with no error. The gray shaded region shows a ±20% error margin, providing a visual reference for acceptable prediction deviations. The right plot shows a parity plot of the simulated vs experimental surface tension at 25 °C for the HM only.

Table 2 shows the HM metrics, both for training and test set. The training set metrics highlight the model accuracy on known data, indicating excellent fit. For the test set, the metrics remain robust, also demonstrating good agreement with experimental data.

Table 2: HM metrics on training and test sets.

Metric	Train	Test
MAE	0.60 mN/m	1.4619 mN/m
MAPE	2.27%	3.77%
MSE	0.6281 mN^2/m^2	3.4852 mN^2/m^2
RMSE	0.7925 mN/m	1.8669 mN/m
R^2	0.99	0.90

SHAP analysis

Figure 4 shows a SHAP violin plot for the presented model. In the violin plot shown, SHAP values are plotted for the features evaluated on the test set, with the horizontal axis representing the SHAP value (i.e., the impact on the model's output) and the vertical axis listing the features. The shape of each violin represents the distribution of SHAP values for a feature across the dataset, highlighting variability in its impact.

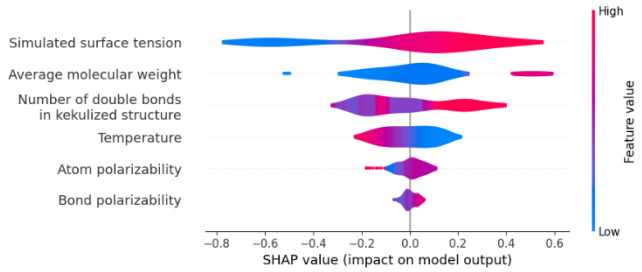

Figure 4: SHAP analysis violin plot. Simulated surface tension is the most relevant feature, bond polarizability the least relevant one.

Additionally, the color gradient within each violin indicates the corresponding feature value, with red denoting high feature values and blue denoting low feature values. This dual representation allows for an intuitive understanding of both the direction and magnitude of feature contributions to the model predictions. For the developed HM, the simulated surface tension is the most influential feature, with higher simulated surface tension (red) generally increasing the model's predictions and lower values (blue) reducing them. The widespread indicates a substantial variability in its impact. The average molecular weight of a molecule is nothing else than the sum of molecular weights of its elements divided by the number of elements. It is strictly correlated with the molecular weight, which can affect the hydrophobicity of a molecule, thus affecting how it behaves at a liquid interface. However, the effect can vary depending on the compound's structure and functional groups. For example, if the molecule is large but contains hydrophobic regions, it may reduce surface tension by disrupting water's hydrogen bonding network. Conversely, molecules with hydrophilic groups and higher molecular weights may increase surface tension by reinforcing interfacial interactions. The average molecular weight feature shows a less pronounced but consistent positive influence on predictions when its values are high, while lower values have little to no effect. Aromaticity can be defined in a multitude of ways. Moreover, 'aromatic' moieties in structures can be represented in Kekulé form using alternating single and double bonds [11]. Double bonds introduce rigidity and delocalized π-electron systems, which can affect molecular interactions. Molecules with a higher number of double bonds in the Kekulized structure could exhibit increased molecular polarizability, which can contribute to higher surface tension. Additionally, these aromatic systems may alter the hydrophobicity or hydrophilicity of a molecule, further affecting how it behaves at liquid interfaces. The number of double bonds in kekulized structure feature has a balanced influence, with higher values slightly increasing the predictions and lower values reducing them. The spread of SHAP values suggests moderate variability in its impact. High temperatures tend to lower the prediction values, while lower temperatures slightly increase them. This behavior aligns with expectations. As temperature rises, the self-solvation energy slightly increases (e.g. becomes less negative). Since surface tension is calculated as the negative ratio of the self-solvation energy to the molar surface (which is temperature-independent), an increase in temperature results in a decrease in surface tension. Polarizability measures how easily the electron cloud around an atom can be distorted by external electric fields. It affects the strength of van der Waals forces, which play a significant role in determining surface tension. Atomic polarizability is a parameter depending on the ease of deforming the valence electron clouds of a chemical species. Mordred iterates over all atoms and bonds of a molecule and then return the sum of all atomic and bond polarizabilities as a total value. Molecules with higher atom polarizability often exhibit stronger intermolecular attractions, potentially increasing surface tension. Similar to atom polarizability, bond polarizability refers to the distortion of electron density within bonds. It influences the molecule's ability to interact with neighboring molecules. While its direct impact on surface tension may be weaker compared to atom polarizability, it can contribute to intermolecular forces. The variability indicates a modest but significant influence. Atom and bond polarizability exhibit a narrower range of SHAP values, indicating a limited impact overall.

CONCLUSIONS

A HM to accurately predict surface tension of organic molecules from 5 to 50°C has been presented. This study highlights the value of integrating first-principle and data-driven methods to enhance prediction. Nevertheless, the developed HM was trained on a relatively small dataset consisting of 68 molecules. The authors of this work have parallelly revisited a comprehensive but largely inaccessible, due to its outdated digital format, dataset of over 2000 surface tension measurements, transforming it into a machine-readable format to eliminate manual transcription errors. The experimental data utilized in this work originates from Jasper's compilation [12]. The dataset will be used for model training in future work. Moreover, the authors are currently working at an improved, physics-informed, version of the current model. In order to prove previous assumption on deviation source, 53 pseudo-experimental self-solvation free energies have been found from dGsolvDB1 dataset [13]. Through the FPM, calculations of the real molecular surface can be performed. The ANN can be therefore trained to also predict the deviation between the flat surface and the real one, in order to show how restrictive this hypothesis can be. Additionally, developing a model capable of pointing the sources of FPM deviations and addressing them could be highly beneficial. Such a model could then be adapted and extended to a wider range of applications.

ACKNOWLEDGEMENTS

FE, UDC, MEL acknowledge funding from the KU Leuven project HyPro (C3/23/007) and VLAIO project FAST (HBC2021.0974).

REFERENCES

1. A. Roosta, P. Setoodeh, and A. Jahanmiri, "Artificial neural network modeling of surface tension for pure organic compounds," *Ind Eng Chem Res*, vol. 51, no. 1, pp. 561–566, Jan. 2012, doi: 10.1021/ie2017459.
2. V. Goussard et al., "Predicting the Surface Tension of Liquids: Comparison of Four Modeling Approaches and Application to Cosmetic Oils," *J Chem Inf Model*, vol. 57, no. 12, pp. 2986–2995, Dec. 2017, doi: 10.1021/acs.jcim.7b00512.
3. E. J. Delgado and G. A. Diaz, "A molecular structure based model for predicting surface tension of organic compounds," *SAR QSAR Environ Res*, vol. 17, no. 5, pp. 483–496, Oct. 2006, doi: 10.1080/10629360600933913.
4. Y. Lei, Y. Shu, X. Liu, X. Liu, X. Wu, and Y. Chen, "Predictive modeling on the surface tension and viscosity of ionic liquid-organic solvent mixtures via machine learning," *J Taiwan Inst Chem Eng*, vol. 151, Oct. 2023, doi: 10.1016/j.jtice.2023.105140.
5. A. Klamt, "Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena," *J. Phys. Chem*, vol. 99, pp. 2224–2235, 1995, doi: 10.1021/j100007a062.
6. C. Nieto-Draghi et al., "A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes," *Chem Rev*, vol. 115, no. 24, pp. 13093–13164, Dec. 2015, doi: 10.1021/acs.chemrev.5b00215.
7. T. Gaudin, "Exploring direct prediction of surface tension from COSMO-RS theory," *Chem Phys Lett*, vol. 706, pp. 308–310, Aug. 2018, doi:10.1016/j.cplett.2018.06.036.
8. J. Glassey and M. von Stosch, Hybrid Modeling in Process Industries, First edition. New York: CRC Press, 2018.
9. H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi, "Mordred: A molecular descriptor calculator," *J Cheminform*, vol. 10, no. 1, Feb. 2018, doi: 10.1186/s13321-018-0258-y.
10. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017.
11. S. J. Cyvln and I. Gutman, "Kekulé Structures and their Symmetry Properties," *Comp. & Maths. with Appls*, vol. 12, pp. 859–876, 1986.
12. J. J. Jasper, "The Surface Tension of Pure Liquid Compounds," *J Phys Chem Ref Data*, vol. 1, no. 4, pp. 841–1010, 1972, doi: 10.1063/1.3253106.
13. Y. Chung, F. H. Vermeire, H. Wu, P. J. Walker, M. H. Abraham, and W. H. Green, "Group Contribution and Machine Learning Approaches to Predict Abraham Solute Parameters, Solvation Free Energy, and Solvation Enthalpy," *J Chem Inf Model*, vol. 62, no. 3, pp. 433–446, Feb. 2022, doi: 10.1021/acs.jcim.1c01103.