ACS ES&T | Air

pubs.acs.org/estair

Article

A Machine Learning Approach for Predicting the Pure-Component Surface Tension of Atmospherically Relevant Organic Compounds

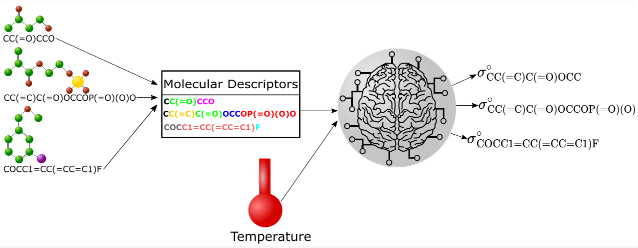
Published as part of ACS ES&T Air special issue "John H. Seinfeld Festschrift".

Ryan Schmedding, Mees Franssen, and Andreas Zuend*

Cite This: *ACS EST Air* 2025, 2, 808−823

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Atmospheric aerosols are complex mixtures of highly functionalized organic compounds, water, inorganic electrolytes, metals, and carbonaceous species. The surface properties of atmospheric aerosol particles can influence several of their chemical and physical impacts, including their hygroscopic growth, aerosol−cloud interactions, and heterogeneous chemical reactions. The effects of the various compounds within a particle on its surface tension depend in part on the pure-component surface tensions. For many of the myriad of organic compounds of interest, experimental pure-component surface tension data at tropospheric temperatures are lacking, thus, requiring the development and application of property estimation methods. In this work, a compiled database of experimental pure-component surface tension data, covering a wide range of organic compound classes and temperatures, is used to train four different types of machine learning models to predict the temperature-dependent pure-component surface tensions of atmospherically relevant organic compounds. The trained models process input information about the temperature and the molecular structure of an organic compound, initially in the form of a Simplified Molecular Input Line Entry System (SMILES) string, to enable predictions. Our quantitative model assessment shows that extreme gradient-boosted descent along with Molecular ACCess System (MACCS) key descriptors of molecular structure provided the best balance of derived input complexity and model performance, resulting in a root-mean-square error (RMSE) of ∼1 mJ m^{-2} in pure-component surface tension. Additionally, a simplified model based on molar mass, elemental ratios, and temperature as inputs was developed for use in applications for which molecular structure information is incomplete (RMSE of ∼2 mJ m^{-2}). We demonstrate that including predicted pure-component surface tension values in thermodynamically rigorous bulk−surface partitioning calculations may substantially modify the critical supersaturations necessary for aerosol activation into cloud droplets.

KEYWORDS: *machine learning, surface tension, qspr, molecular descriptors, aerosols*

■ INTRODUCTION

Atmospheric aerosols are suspensions of particles and the gas phase that surrounds them. Atmospheric aerosol particles can modify the global climate both directly by scattering and reflecting incoming and outgoing solar radiation and indirectly through their impacts on clouds by acting as cloud condensation nuclei (CCN) for liquid droplets or ice nucleating particles (INP) in ice and mixed-phase clouds.[1,2] For complete lists of abbreviations, symbols, and their meanings, please refer to Tables S1 and S2 in the Supporting Information. The activation of aerosol particles into cloud droplets is governed by several factors. The critical supersaturation necessary for cloud droplet activation is given by the global maximum of the following equation:[3]

$$S = a_w \exp\left(\frac{4\sigma M_w}{RT\rho_w D_p}\right) \tag{1}$$

Here, S represents the equilibrium saturation ratio, a_w is the activity of water in the particle, and M_w and ρ_w are the molar mass and density of water, respectively. R is the gas constant, T is the temperature, D_p is the aerosol particle diameter, and σ is the equilibrium droplet surface tension. This equation can be broken up into the Raoult (or solute) effect which is described by a_w and the Kelvin (or surface effect) which is represented by the exponential factor in eq 1.

Recently, the role of surface properties and corresponding feedbacks on the Raoult effect in the activation of aerosol particles into cloud droplets has come under scrutiny.[4−14]

Received: October 16, 2024
Revised: February 18, 2025
Accepted: February 19, 2025
Published: April 8, 2025

https://doi.org/10.1021/acsestair.4c00291

Beyond cloud droplet activation, surface tension is crucial in determining the size-dependent surface composition of aqueous aerosol particles.[14] The composition of the surface of aerosol particles may influence multiphase chemical reactions at/near the surface,[15,16] ice nucleation capability,[17] optical properties,[18] the presence or suppression of liquid–liquid phase separation (LLPS),[19] the transport of contaminants such as per- and polyfluoroalkyl substances (PFAS),[20] the hygroscopic growth of nanometer-scale glassy aerosol particles,[21] and several other surface-influenced processes. For a thorough review of the importance of aerosol surface properties, we refer to Wokosin et al.[22]

Because of the aforementioned importance of aerosol surface properties, various models have been developed that account for surface tension (σ) modifications as a function of aerosol composition and coupled changes in bulk–surface partitioning in aerosol particles.[14,23−25] Surface tension can be thought of physically as the work per unit surface area required to expand a 2-dimensional phase boundary. On a molecular, chemical level, surface tension can be interpreted as the additional energetic penalty that a molecule at a gas–liquid interface experiences when it is unable to interact with other closely spaced liquid-phase molecules around it in all directions.[26] Additionally, in the context of LLPS particles, compositional differences between a surface phase and an underlying, organic-rich bulk phase may be reduced compared to a single-phase aqueous particle state.[19] Due to nonideal mixing in aqueous phases among the various mixture species, organic molecules composed of distinct amounts of hydrophilic and/or hydrophobic functional groups typically partition between a droplet's bulk and surface differently, the equilibrium state of such nonideal mixing effects can be predicted by thermodynamic models.[14,27,28] In the case of surfactants, surface-enrichment is prevalent even in the presence of LLPS, and this surface enrichment is typically enhanced with decreasing droplet size and corresponding increasing surface area to volume ratio.[7,11,25,29−34]

Many of the models for predicting the effective surface tension of aerosol particles and cloud droplets rely on accurate representations of pure-component surface tension values (σ_i°) at temperatures of interest; hereafter, we use superscript ° to denote a pure-component rather than mixture property. Models for the surface tension of multicomponent solutions, which rely on accurate values of pure-component surface tension (σ_i°), range in complexity. The semiempirical Eberhard model for binary solutions relates the bulk mole fractions of species j and k (x_j and x_k) and their corresponding pure-component surface tensions (σ_j° and σ_k°) to the solution surface tension through a fitted parameter (s_{jk}) as follows:[35]

$$\sigma = \frac{x_j\sigma_j^\circ + s_{jk}x_k\sigma_k^\circ}{x_j + s_{jk}x_k} \tag{2}$$

A more complex model for binary solutions was derived by Connors and Wright[36] and later by Shardt and Elliott[37]

$$\sigma = \sigma_k^\circ - \left(1 - \frac{b(1-x_j)}{1-a(1-x_j)}\right)x_j(\sigma_j^\circ - \sigma_k^\circ) \tag{3}$$

where a and b are semiempirical parameters. This approach was later shown to be applicable to multicomponent solutions by Shardt et al.[38] and an extension of eq 2 by Kleinheins et al.[29] The Sprow–Prausnitz–Butler equation is a thermodynamically rigorous treatment of solution surface tension that also relies on accurate σ_i° values:[39]

$$\sigma = \sigma_i^\circ + RT\ \ln\left(\frac{a_i^{\mathrm{s}}}{a_i^{\mathrm{b}}}\right) \tag{4}$$

In eq 4, R is the ideal gas constant, T is the absolute temperature, a_i^{s} is the (chemical) activity of i in the surface phase, and a_i^{b} is the activity of i in the bulk phase. Beyond predicting the value of σ for solutions, accurate values for σ_i° may also be necessary for predicting interfacial tensions between two liquids, α and β, should LLPS occur in an aerosol particle.[19] Approaches for interfacial tension estimation include Antonov's rule[40]

$$\sigma^{\alpha\beta} = |\sigma^\alpha - \sigma^\beta| \tag{5}$$

and the Girifalco–Good equation[41]

$$\sigma^{\alpha\beta} = \sigma^\alpha + \sigma^\beta - 2\phi\sqrt{\sigma^\alpha\sigma^\beta} \tag{6}$$

where $\sigma^{\alpha\beta}$ is the interfacial tension between phases α and β, and ϕ is a fitted parameter that is often assumed to be 1.0. The values of σ^α and σ^β may be calculated using one of the above models for solution surface tension or other simplified mixing-rule-based models which depend on σ_i°.[19]

The importance of accurately describing σ_i° is evident. However, because of the high complexity and degree of functionality of many of the myriad of organic aerosol compounds, a rather limited set of measurements for σ_i° of organic species exists.[23,30,42−44] As such, predictive models or simplified assumptions about the surface tensions of organic compounds must be employed in many applications.

The simplest treatment of atmospherically relevant organic species is to assume that all have the same value.[9,12,14] The unknown pure-component surface tension of species has also been estimated based on measured values of chemically similar species or by extrapolating the measured behavior of highly concentrated binary aqueous solutions of organic species toward the pure-component limit.[14,19] Shardt and Elliott[37] estimated σ_i° by fitting a simple linear equation with two parameters ($\theta_{0,i}$ and $\theta_{1,i}$) to include a temperature dependence:

$$\sigma_i^\circ = \theta_{0,i} + \theta_{1,i}T \tag{7}$$

This approach requires detailed temperature-dependent data for σ_i° and was only used for 15 organic compounds in Shardt and Elliott.[37]

Another semiempirical approach to predicting the surface tension of pure compounds is through the use of the Macleod–Sudgeon parachor, which is defined as follows:[45,46]

$$\sigma_i^\circ = [\mathcal{P}_i(T)\cdot(\rho_i^{\mathrm{l}} - \rho_i^{\mathrm{v}})]^4 \tag{8}$$

The parachor, $\mathcal{P}_i$, is a semiempirical term relating the difference between the liquid-state density (ρ_i^{l}) and the vapor-state density (ρ_i^{v}) to the pure-component surface tension at a given temperature (T). Owing to the simplicity of the parachor approach, it has been used extensively in fields outside of atmospheric science to predict the surface tension of organic compounds.[47−49] The parachor method can also be modified to predict the surface tension of solutions;[50,51] however, it has been noted that such modifications perform poorly for solutions of water and organic compounds.[47] Indeed, the functional form of $\mathcal{P}_i$ is poorly constrained, and various methods have been proposed to describe it as a weak function of temperature.[45,46,52] Escobedo and Mansoori[53] related the value of $\mathcal{P}_i$ to a function of the reduced temperature (T_r), $T_r = \frac{T}{T_c}$, through the following equation:

$$\mathcal{P} = \mathcal{P}_\circ(1 - T_r)^{0.37}\cdot T_r\cdot\exp(0.30066/T_r + 0.86442T_r^9) \tag{9}$$

$\mathcal{P}_\circ$ was defined in the same work as follows:

$$\mathcal{P}_\circ = 39.643\left(0.22217 - 2.91042\times10^{-3}\frac{\mathcal{R}^\star}{T_{b,r}^2}\right)T_c^{13/12}P_c^{5/6} \tag{10}$$

$\mathcal{R}^\star$ is the ratio of the molar refractivity of compound i to the molar refractivity of i in methane. $T_{b,r}$ is the reduced boiling point of i, T_c is the critical temperature, and P_c is the critical pressure of i. Escobedo and Mansoori[53] found that such equations were able to predict the pure-component surface tension of 94 different compounds to within 2.5% absolute average percent deviation. However, it should be noted that such an approach requires knowledge of numerous physicochemical properties of individual components in order to compute $\mathcal{P}_i$ and by extension σ_i°.

Other approaches to calculate σ_i° include Density Functional Theory (DFT).[54,55] DFT relates the surface tension of i to the difference of the Grand Potential (Ω) in the surface and in the bulk phase as follows:

$$\sigma_i^\circ = \frac{1}{A}(\Omega^{\mathrm{s}} - \Omega^{\mathrm{b}}) \tag{11}$$

A is the area of the surface in this case. In order to calculate Ω^{s} and Ω^{b}, the Helmholtz energy and the chemical potential of species i must be known.[56] Using density functional theory to predict σ_i° has several limitations. DFT calculations are computationally expensive and require numerous assumptions about the structure of the surface region. DFT calculations may provide information about the orientation and density of molecules in the surface region; however, they are not typically used in applications in the field of aerosol science due to the aforementioned computational limitations.

Another method of predicting σ_i° at a given temperature is the Theory of Corresponding States,[57] which relates the reduced surface tension of a compound to the critical pressure and temperature of the compound:

$$\sigma_i^\circ(T) = \sigma_i^{\mathrm{ref}}\left(1 - \frac{T}{T_c}\right)^{n_i} \tag{12}$$

Here, σ_i^{ref} is a constant reference surface tension, and n is an empirical (fit) coefficient, equal to $\frac{11}{9}$ in the ideal case but it may range between 1.16 and 1.5 in real systems.[57,58] T is the absolute temperature of the system in K, and T_c is the critical temperature of i in K.

Numerous empirical parametrizations besides the aforementioned method of Escobedo and Mansoori[53] have been developed based on the critical properties of compounds such as the approach of Brock and Bird:[59]

$$\sigma_i^\circ = P_c^{\frac{2}{3}}T_c^{\frac{1}{3}}Q(1 - T_r)^{\frac{11}{9}} \tag{13}$$

Here, the factor Q can be defined as follows:

$$Q = 0.1196\left[1 + \frac{T_{b,r}\ln(P_c/101325)}{1 - T_{b,r}}\right] - 0.279 \tag{14}$$

In eq 13 and eq 14, P_c has units of Pa, and T_c and T_b have units of K, such that σ_i° has units of J m^{-2}. Other empirical approaches include additional input parameters, which must be known for predicting σ_i°:[60]

$$\sigma_i^\circ = \varphi\frac{M_i^{1/3}}{6N_A^{1/3}}\rho_l^{2/3}[H_{\mathrm{vap},T_b} + C_{p,l}\cdot(T_b - T)] \tag{15}$$

In eq 15, M_i is the molar mass (kg mol^{-1}), N_A denotes Avogadro's number, and ρ_l is the liquid-state density (kg m^{-3}) at T and P. H_{vap,T_b} is the enthalpy of vaporization at T_b in units of J kg^{-1}, $C_{p,l}$ is the liquid-state heat capacity at constant pressure at T in units of J kg^{-1} K^{-1}, and φ can be defined as follows

$$\varphi = 1 - 0.0047M_i + 6.8\times10^{-6}M_i^2 \tag{16}$$

such that σ_i°, as calculated by eq 15, has units of J m^{-2}. Gharagheizi et al.[61] proposed another empirical equation which relies on fewer input parameters

$$\sigma_i^\circ = 8.948226\times10^{-4}\left[\frac{A^2}{M_i}\sqrt{\frac{A\omega}{M_i}}\right]^{1/2} \tag{17}$$

with A defined as

$$A = T_c - T - \omega \tag{18}$$

where ω denotes the acentric factor of component i. σ_i° carries units of J m^{-2}.

We note that many of these empirical equations were not trained or evaluated on diverse classes of input data. For example, eq 15 was only fitted using alkanes of chain lengths from C_1 to C_{10} and C_{12}, and thus, its utility for functionalized organic compounds may be limited.[60] More recent attempts to model the surface tension of functionalized organic species have relied on statistical regression methods[62,63] or artificial neural networks (ANN).[64−72]

Artificial neural networks loosely mimic the activity of a brain by containing mathematical representations of neurons grouped into layers. Each artificial neuron contains an activation function that takes inputs and returns the value of said function similar to a biological neuron's action potential in the brain of an animal. Artificial neurons are then grouped into sequences of layers. The artificial neurons in each layer are connected to both the previous and subsequent layers of artificial neurons such that the outputs of the previous layer become the inputs of the current layer, with the final layer producing the output of interest. As the activation functions for each neuron in each layer may return different values, the input values are transformed by passing them through multiple layers of neurons until an acceptable result is produced. In a regression problem where a single value is desired, in this case, surface tension, a final single neuron is used to generate the output. The number of inner layers and the number of neurons per layer in this type of artificial neural network must be found through a trial and error procedure to avoid underfitting or overfitting of the model. A brief summary of different statistical and machine learning (ML) modeling approaches for estimating the surface tension of organic compounds follows.

Sanjuán et al.[62] used surface tension values for 87 different alcohols and compared models based on various combinations of temperature, triple point, normal boiling point, and critical temperatures; triple point and critical pressures, critical compressibility factor, critical volume, molar volume, molar mass, radius of gyration, and acentric factor. They found that models that depended on temperature, critical temperature, critical pressure, critical volume, molar volume, and acentric factor had the best correlations with the measurements in question.[62] Roosta et al.[73] used an ANN with a single hidden layer and 20 nodes and was able to accurately predict the surface tensions of organic compounds across a broader temperature range than Escobedo and Mansoori,[53] with inputs based on a component's critical pressure, acentric factor, reduced normal and boiling temperature, and specific gravity at the compounds normal boiling point temperature. Randová and Bartovská[63] used a group contribution method and eq 12 to predict the surface tension of straight-chain and branched alkanes.

Lazzús et al.[68] used 46 different functional groups, the molar mass of the compound, and the absolute temperature to predict the surface tension of different ionic liquid compounds using 1 hidden layer with 30 neurons followed by a gravitational search algorithm to predict σ_i°. Lashkarbolooki and Bayat[69] specifically examined the surface tension of alkanes and alkenes using an ANN with 1 hidden layer with 27 neurons that took absolute temperature, critical temperature, and number of carbons as inputs. Mousavi et al.[67] also examined a functional-group-based approach for determining the surface tension of ionic liquids using a combination of a firefly algorithm and the differential evolution method to optimize a radial-basis function model which takes chemical structure and temperature as inputs. Pierantozzi et al.[70] used a single hidden layer with 41 neurons which took reduced temperature, boiling temperature, and acentric factor for organic acid species to predict their surface tension. Another common application of ANNs is in image recognition and the related evaluation of graphical measurement data. For example, Soori et al.[64] was able to use an ANN to predict the surface tension of binary solutions based on images taken during pendant drop experiments.

Other types of ML-based approaches have been used to predict molecular properties.[74] A commonly used alternative to ANNs is tree-based models. These models have recently been confirmed to perform better for prediction problems when inputs are tabular, and it is simpler to optimize a tree-based model's learning process compared to other ML-based approaches.[75] The simplest tree-based models are decision trees, which utilize a sequence of branching nodes to classify the inputs and predict a value. While individual decision trees are easy to interpret, they are prone to overfitting and may perform poorly for compounds outside of the training database.[76] Tree-based ensemble methods have also been developed as a way to create more powerful models. Two such ensemble models are random forests and gradient-boosted trees. Both techniques utilize multiple decision trees in their model architectures to generate more robust predictions. Random forests generate many decision trees, each of which is trained from a small subset of the overall data set and takes the average of the predictions as a final result. In comparison, gradient-boosted trees create many decision trees in sequence, where each sequential tree is trained on the residuals of the previous tree's predictions. One of the more popular variants of gradient boosted trees is the extreme gradient boosted descent (XGB). XGB has been widely used and generally has been found to perform well for predicting molecular properties.[77] An additional feature that is unique to an XGB model is that it allows for explicit monotonic constraints on the relationship between input values and model predictions. Such a parametrization allows for XGB models to more easily represent physical behavior in a realistic manner, such as the inverse relationship between T and σ_i° wherein an increase in T leads to a decrease in σ_i°.

While a combination of various ML techniques and functional group-based approaches may have a high degree of flexibility for many quantitative structure−property relationship (QSPR) approaches, such approaches may not be able to adequately represent numerous compounds to a high degree of accuracy. For example, cis−trans isomerism may lead to substantial differences in pure-component surface tension.[78] Thus, more thorough methods are desirable to characterize individual molecules from their simplified molecular input line entry system (SMILES) notation, which is a method of representing molecules as a single string of characters. One such technique is known as molecular fingerprinting; it involves translating a molecular structure into a series of integer codes that represent the molecule's structure. Two of the more common methods of molecular fingerprinting are Molecular ACCess System (MACCS) keys and so-called Morgan fingerprints, both of which are widely used in the field of cheminformatics to predict QSPR for different molecular properties.[79]

MACCS keys are a set of 166 predefined patterns that can be present in a molecule.[80] Each of these patterns is associated with a corresponding SMARTS (SMiles ARbitrary Target Specification) code. The use of SMARTS allows for the parsing of a given SMILES string to count the matching patterns. Such a defined list of MACCS keys allows for shorter descriptions of molecular features and thus lower computational costs when used as inputs in QSPR models. Because SMARTS codes are designed to operate on SMILES, MACCS keys are interpretable by human readers since all that is necessary is a reference table with a MACCS key number and the corresponding SMARTS code and molecular pattern. There are some disadvantages to using MACCS keys. For example, due to the limited number of patterns that MACCS keys describe, they may not be able to accurately represent more complex and nuanced patterns in a large, multifunctional molecule. One example of this limitation is MACCS key 44, which is simply labeled as "other" and describes any feature not captured by the remaining 165 keys.

In comparison, Morgan fingerprints can encode much more information about an individual molecule's structure, albeit in a more abstract way. Morgan fingerprints are in the family of extended-connectivity fingerprints, which numerically encode each atom in a molecule and the local structure around said atom within a given radius of adjacent atoms, typically two.[81] Recently, Orsi and Reymond[82] developed a novel fingerprinting technique that modifies hashed Morgan fingerprints to include data on the chirality of each atom in a molecule, thus increasing the level of structural detail.

METHODS

Data Collection and Processing. Surface tension data were collected for 1805 unique organic species reported by Jasper[83] on 274 original references. Measurements were principally taken using the capillary rise method against the ambient atmosphere. Additional measurements were reported using the capillary rise method against N_2 or a vacuum, the differential capillary rise method against the ambient atmosphere or N_2, the maximum bubble pressure method against the ambient atmosphere or N_2, the drop weight method against the ambient atmosphere or N_2, the sessile drop method, horizontal capillary method against the atmosphere, and the ring detachment method against the ambient atmosphere. It should also be noted that 61 of the unique compounds did not have a measurement method reported by Jasper.[83] In the aforementioned work, two T-dependent parameters were fit to measured σ_i° values to predict σ_i° within 0.2−0.3 mJ m^{-2} and used to generate additional σ_i° values as a function of T. With the addition of T as a parameter, there were 12446 σ_i° and T data pairs reported by Jasper,[83] which were suitable for use in this work. Stereoisomers were also counted as two distinct compounds, where measurements were available and sufficiently specific. Compounds with only one σ_i° and T data pair were also included to increase the size of the training data. However, the number of compounds with surface tension measurements at a single T was low; of the 12446 σ_i° values, only 164 were limited to surface tension data at a single T. All of the compounds for which a measurement technique was not reported had a single σ_i°-T pair reported. Compounds included in the data set contained the following elements: carbon (1805 distinct compounds), hydrogen (1763 distinct compounds), oxygen (1143 distinct compounds), nitrogen (341 distinct compounds), sulfur (130 distinct compounds), phosphorus (73 distinct compounds), fluorine (78 distinct compounds), chlorine (148 distinct compounds), bromine (58 distinct compounds), and iodine (30 distinct compounds).

Figure 1 shows a 2D kernel density estimate of both σ_i° and temperature for all compounds in the training and test databases. Figure S1 shows the kernel density estimate for groups of compounds with more than three unique SMILES that contain only one type of nonhydrocarbon functional group, for example, compounds only containing carboxyl or amide groups. Figure S2 corresponds to Figure S1 but for compounds with multiple nonhydrocarbon functional groups; for example, all compounds that contain a hydroxyl group and any other nonhydrocarbon functional group. Reported T values in the complete set of data ranged from 113 to 523 K, and σ_i° ranged from 8.4 to 68.8 mJ m^{-2}. For temperatures between 218 and 318 K, which corresponds to a typical range in the troposphere, there are 5557 σ_i° values, which correspond to 44.7% of the total data set. The majority of the remaining data correspond to temperatures above 318 K. From the data set described in the preceding paragraphs, isomeric SMILES were generated using OEChem v2.3.0 through PubChem release v2021.05.07.[84] A link to the complete database of compounds and their SMILES and $\sigma_i^\circ-T$ pairs can be found in the Data Availability section at the end of this work.

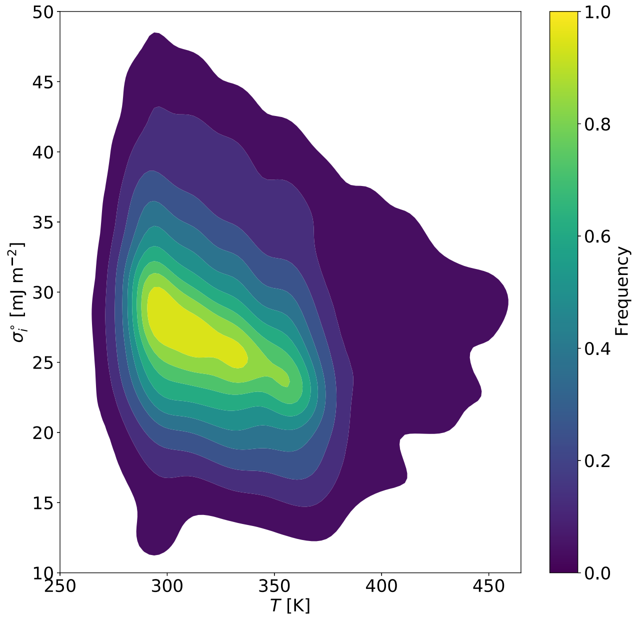


Figure 1. A two-dimensional kernel density estimate of σ_i° of all 1805 unique compounds and the corresponding temperatures for which σ_i° values were reported. The median and mean σ_i° values are 26.2 mJ m^{-2} and 27.3 mJ m^{-2}, respectively. The median and mean temperatures are 323 K and 329 K, respectively.

Model Architecture. It is important to note that the overall architecture of a machine-learning model, rather than the individual weights of a model, is itself tunable and impacts the model performance. Tunable model features that may constrain model weights and overall performance are known as hyperparameters and also include algorithmic features such as the learning rate and batch size. In order to predict temperature-dependent σ_i° values, various models and their hyperparameters were optimized and tested. It is likewise useful to determine which categories of models produce the same or similar results if their hyperparameters and weights are optimized. Thus, an extreme gradient boosted descent (XGB), a decision tree (DT), a random forest (RF), and K-nearest neighbors (KNN) regression models were also tested. Prior to training the models, 10% of the σ_i° and T data pairs were randomly selected and set aside for testing following model selection, hyperparameter tuning, and model training. For consistency, the selected test data were kept the same for all of the model types and architectures tested in this work.

A major concern that is often encountered when training ML models is that of an overfitted model, i.e., one that is only capable of reproducing the training data reliably. In the event that a model is overfit, it may perform poorly when novel inputs (here, molecular structures of unseen compounds) are introduced. One method to reduce the likelihood of overfitting a model is through the use of k-fold cross-validation in the training stage. k-Fold cross-validation is the process of splitting up a training data set into k slices, typically five or ten slices,[85] and then training the model on $k - 1$ slices and validating the model performance on the remaining slice. A model performance value is calculated for each slice used as validation data; i.e., all permutations for a slice being the "remaining slice" are run; the mean performance value is used to assess the overall model performance. Typical performance metrics include the root-mean-square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

the mean square error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

the mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{21}$$

and the mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{22}$$

In eqs 19, 20, 21, and 22, n is the number of points in the sample, y_i is the original (known) value, and $\hat{y}_i$ is the predicted value from the model. Another model performance metric is the coefficient of determination, R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{23}$$

In this work, MSE was selected as the primary model performance metric, as it is more sensitive to predictions with large errors and thus may lead to a model that performs well for many different inputs.

An additional challenge in ML-based approaches for predicting molecular properties is determining the optimal model parameters for a given model type in order to maximize the model's performance. Past methods for selecting the best combination of model parameters involved grid searches or random searches of the parameter combinations. While robust, such approaches can prove computationally expensive to thoroughly explore all possible parameter combinations. Bayesian optimization algorithms have been found to perform quite well in a fraction of the time that traditional grid or random searches require.[86] The hyperparameters for all models in this work were optimized using the Optuna v3.6.1[87] Bayesian optimization algorithm. All hyperparameter tuning, model training, cross-validation, testing, and plotting were performed using the CryoCloud JupyterBook.[88]

Because both MACCS keys and Morgan fingerprints can provide useful information about a molecule, a set of the above models was trained by using either MACCS keys or Morgan fingerprints along with temperature as inputs. MACCS keys and Morgan fingerprints were both generated from a component's SMILES using RDKit v2024.3.5[89] in Python v3.11.9. XGB models were developed using the XGBoost v2.1.0,[90] and the RF, DT, and KNN models were generated using SciKit-learn v1.5.1.[91]

For situations wherein complete molecular structure information may not be available, a simplified XGB model was also developed that took the following molecular properties as inputs: temperature (T), molar mass (M_i), and the following atomic ratios: O:C, H:C, N:C, S:C, P:C, Cl:C, F:C, I:C, and Br:C. These inputs were selected based on their likely availability from field observations, such as from aerosol mass spectrometer measurements, as well as values that could be easily calculated for surrogate compounds used to represent the various organic aerosol components in chemical transport models.

Model Testing. Given the three models introduced in this work, a comparison to past methods of predicting σ_i° was carried out by utilizing the critical properties found in Yaws[92] along with eq 13 and eq 17 to predict σ_i° for shared compounds between those reported by Yaws[92] and those reported by Jasper.[83] eq 13 and eq 17 were specifically selected because they relied on the fewest inputs among the empirical relationships discussed in the Introduction. Matched compounds were specifically selected out of the test data set that had been previously set aside from the training data to avoid any possible artifacts from model training.

In order to determine if there were additional features that may influence surface tension that were not accounted for by any of the models, the 100 poorest performing σ_i° values were extracted from the best performing model for each of the three types of model inputs. In the case of the simplified inputs, a Student's t test and Brunner-Munzel test were performed for each of the model inputs, including temperature, to determine whether there may have been a significant difference between these compounds and the remainder of the test data set. For both the MACCS keys and the Morgan fingerprints, a Student's t test and Brunner-Munzel test were once again performed for temperature in comparison to the remainder of the test data. However, to better understand any structural artifacts that may not have been captured by the model, Tanimoto similarity values were calculated for all combinations of compounds. Tanimoto similarity ($S_{j,k}$) is a measure of the structural similarity between two compounds

$$S_{j,k} = \frac{j \cap k}{j \cup k} \tag{24}$$

where j and k are the sets of fingerprints or descriptors that represent molecule j and molecule k, respectively.[93] In the case of two identical molecules, the Tanimoto similarity score is equal to 1, and in the case of two molecules that do not share any overlapping features, the Tanimoto score is 0. Following the calculation of Tanimoto scores for all possible pairs of compounds, a similarity matrix can be constructed, which can then be used to construct hierarchical clusters of the poorest-performing compounds.

Model Applications. As mentioned in the Introduction, σ_i° is an important parameter for determining the effective solution surface tension for a variety of environmentally relevant systems. One such application is by utilizing the Butler−Sprow−Prausnitz equation (eq 4) to compute the equilibrium surface compositions and tensions of aerosol particles and cloud droplets, following the approach of Schmedding and Zuend.[14] Such an approach relies on calculating the activity coefficients of each species in a droplet using the Aerosol Inorganic−Organic Mixtures Functional groups Activity Coefficients (AIOMFAC) model.[27,28] Because many aerosols are chemically complex mixtures of secondary compounds formed from the emissions of biogenic precursors, two different systems were tested. The first system is comprised of 21 isoprene-derived organic compounds generated by the Master Chemical Mechanism (MCM)[94,95] along with ammonium sulfate (to represent inorganic electrolytes). The second system consisted of 14 α-pinene oxidation products along with ammonium sulfate.[95,96] Because the effect of surface tension reductions becomes most noticeable for particles with diameters less than 100 nm, both systems were run assuming a water-free (dry) particle diameter of 50 nm. The pure-component surface tension of water at 298 K was set to 71.98 mJ m^{-2} for all systems,[83] and the pure-component hypothetical surface tension of (liquid) ammonium sulfate with air was set to 184.5 mJ m^{-2} following the approach of Dutcher et al.,[97] which extrapolates surface tension at lower temperatures from molten salt values for inorganic electrolyte systems. Once the effective droplet surface tension is calculated as a function of droplet composition, a Köhler curve can be calculated using eq 1. The global maximum of S in eq 1 represents the critical saturation ratio that must be reached or exceeded in the air surrounding that specific aerosol particle for

Table 1. Performance Metrics Generated from the 1245 Test Data Set Pairs for Different Input Types (SMILES-Generated MACCS Keys, SMILES-Generated Morgan Fingerprints, and Simplified Elemental Ratios) and Model Types Tested in This Work

Input type	Model type	MSE	RMSE	MAE	MAPE	R^2
Simplified	Extreme Gradient Boosting (XGB)	3.883	1.970	1.329	5.1%	0.905
	K-Nearest Neighbors (KNN)	21.978	4.68	3.031	10.9%	0.463
	Random Forest (RF)	5.537	2.353	1.640	6.2%	0.865
	Decision Tree (DT)	7.342	2.709	1.883	7.0%	0.821
MACCS key	Extreme Gradient Boosting (XGB)	1.156	1.076	0.660	2.7%	0.972
	K-Nearest Neighbors (KNN)	10.226	3.198	2.028	7.6%	0.750
	Random Forest (RF)	1.872	1.368	0.916	3.7%	0.954
	Decision Tree (DT)	2.725	1.651	2.203	4.7%	0.933
Morgan fingerprint	Extreme Gradient Boosting (XGB)	1.012	1.006	0.633	2.5%	0.975
	K-Nearest Neighbors (KNN)	11.472	3.387	2.195	8.2%	0.720
	Random Forest (RF)	3.126	1.768	1.252	5.0%	0.924
	Decision Tree (DT)	4.080	2.020	1.460	5.6%	0.900

the particle to activate and grow into a much larger cloud droplet.

RESULTS

Table 1 lists the values of the MSE, RMSE, MAE, MAPE, and R^2 generated from the test data set. This data set contains 1245 $\sigma_i^\circ - T$ pairs or 10% of the overall number of $\sigma_i^\circ - T$ pairs used in this work. Each model (XGB, RF, DT, and KNN) used the following three input categories: Simplified inputs (Simp), MACCS keys (MACCS), and Morgan fingerprints (MF). The MACCS-XGB and the MF-XGB models had the largest R^2 values of similar magnitude; hence, they are considered the best predictive models for our application. Overall, the MF-XGB performed slightly better than the MACCS-XGB for predicting σ_i° when considering all of the compounds. However, we note that the models with MACCS keys inputs use substantially fewer independent (input) variables than the models with Morgan fingerprint inputs, yet they perform slightly better for three model types, and similarly in the case of XGB, as the Morgan fingerprint approach.

Figure 2 shows the predicted σ_i° values from the Simp-XGB, MACCS-XGB, and MF-XGB models described in Table 1 compared to reported σ_i° values from the test data set selected from.[83] Also shown in Figure 2 are the predictions from the empirical approaches of Gharagheizi et al.[61] (eq 17 and eq 18) and Brock and Bird[59] (eq 13 and eq 14) with critical parameters taken from Yaws.[92] We note that only compounds from the test data with critical parameters included in Yaws[92] are plotted, reducing the number of points shown to 429 from the test data. Figure 2A shows the Simp-XGB model, Figure 2B shows the MACCS-XGB model, and Figure 2C shows the MF-XGB approach. In the case of Simp-XGB, the performance was (expectedly) worse than for both the MACCS-XGB and the MF-XGB models, although the model still performed reasonably well with $R^2 = 0.905$, MSE = 3.883, and RMSE = 1.970 mJ m^{-2}. Both the MACCS-XGB and MF-XGB models generally perform better for compounds with higher σ_i° values than those with lower σ_i°.

The previously described empirical approaches both performed poorly in comparison to all three ML-based approaches shown in Figure 2. The approach of Brock and Bird[59] (eq 13 and eq 14) generally overpredicted σ_i° at higher reported σ_i° values, and the approach of Gharagheizi et al.[61] (eq 17 and eq 18) generally underpredicted σ_i° in comparison to the reference values from Jasper.[83]

For comparison to the more complex models, the importance of the various simplified inputs used with the Simp-XGB model is shown in Figure 3A. Here, importance is quantified in terms of the explained variance of the surface tension prediction. In the model with simplified inputs, T was responsible for 15.4% of the explained variance. Of the simplified inputs representing molecular properties, the molar mass, O:C ratio, H:C ratio, and N:C ratio were the most important and explained collectively 56.5% of the model's variance. Because of the ease of interpreting MACCS keys, the top ten MACCS keys that explained the most variance in the MACCS-XGB model were extracted and are displayed in Figure 3B. The temperature for which the model is run explains the most substantial portion of the variance at 22.7%, and the ten most important MACCS keys explain 71.4% of the MACCS-XGB model variance with those inputs.

The following ten features were MACCS keys corresponding to the following patterns. The most important individual MACCS key is NAO, which represents a substructure in which nitrogen is indirectly connected to an oxygen atom via any intermediate atom. ACH2O represents a methylene group connected to any atom on one side and to an oxygen atom on the other side. These atoms may be connected by any type of bond. ACH2AACH2A represents a more complex pattern, which can be any two atoms between two methylene groups, which themselves are between any other two atoms. In the event that the outer two atoms are the same, this key represents a ring structure containing two methylene groups and two other atoms between them. This key can also represent the same sequence, but if one of the methylene groups is outside of the ring, the substructure becomes one in which a methylene group is connected to any atom and then forms a ring structure with another methylene group and two other ring atoms. CH3AACH2A represents a methyl group connected to any two atoms followed by a methylene group connected to another atom of any type. The bonds between these atoms may be of any type, as well. This key can also represent a ring system where a methyl group is connected to a sequence of atoms that includes a methylene group that is part of the same ring. CH3ACH2A represents a methyl group connected by any bond to any atom that is also connected by any bond to a methylene group. This methylene group is further connected by any bond to another atom. N encodes the presence of nitrogen atoms. A!N$A is a pattern that represents a nitrogen atom that is part of a ring but is also connected to

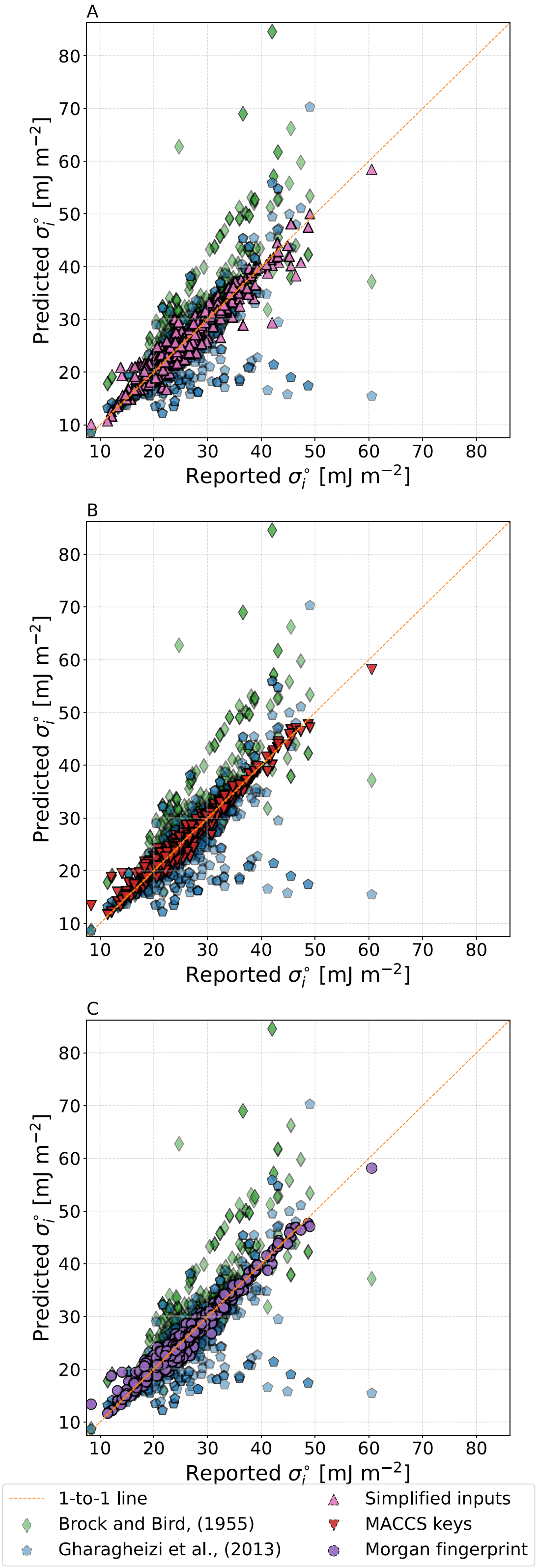


Figure 2. Accuracy of the best-performing model (XGB) in comparison to a subset of the test data for each of the input types: (A) simplified inputs, (B) MACCS keys, and (C) Morgan fingerprints. Also shown are estimations by the empirical methods of Brock and Bird[59] (eq 13) and Gharagheizi et al.[61] (eq 17) for compounds covered by both the test data set and the critical properties data available from Yaws.[92]

another atom outside the ring. A!A$A!A represents a sequence of any four atoms wherein the first and fourth atoms are not a part of a ring system, and the second and third atoms are a part of a ring system. N=A is a structure where nitrogen is connected to any other atom by a double bond. QQ > 1 is a structure where any two atoms that are neither carbon nor hydrogen are connected by any type of bond. Cumulatively, these ten MACCS codes along with T explain 94.1% of the variance in the model.

Figure 3C shows the SMILES for the top ten Morgan fingerprints, the most numerous and complex types of inputs used with the MF-XGB model. For this model, somewhat surprisingly, T was responsible for only 0.2% of the explained variance. Oxygen stands out as the most significant noncarbon element, with SMILES containing at least one oxygen atom accounting for 29.7% of the explained variance. Halogen groups come next, contributing 16.1% to the variance. Among them, fluorinated substructures are the most impactful, explaining 9.8% of the variance, followed by iodine (3.1%), chlorine (1.8%), and bromine, which has the smallest impact among the halogens at 1.6%. Nitrogen and sulfur atoms contribute 8.7% and 5.7% of the variance, respectively. Phosphorus-containing substructures play a minimal role, responsible for just 1.2% of the explained variance. For the less abundant elements like bromine and phosphorus, their impact is likely small due to being present in only a small subset of compounds considered. If temperature is excluded as input for the MF-XGB model, then 186 unique inputs derived from a molecule's structure are required to explain 85% of the variance, and 650 unique inputs derived from a molecule's structure are required to explain 95% of the variance.

To further understand the role of temperature in each of the models, σ_i° predictions were generated for commercially available humic acid (SMILES: C1C2C=CC1C(C2C(=O)O)-(C(=O)O)[N+](=O)[O-]) and commercially available fulvic acid (SMILES: CC1(CC2=C(CO1)C(=O)C3=C(O2)C=C-(C(=C3C(=O)O)O)O)O)O) as a function of the temperature. Both humic and fulvic acids are commonly found in aerosols from biogenic sources, including biomass-burning particles, and can therefore experience a broad range of temperatures from very high values near combustion events to much lower temperatures when they are lofted to higher altitudes in the atmosphere. Therefore, the surface tension of these compounds was studied from 200 to 350 K to include atmospherically relevant temperatures that any aerosol species may experience as well as temperatures that biomass-burning species may experience in a smoke plume. Temperature-dependent surface tensions for commercially available humic acid are shown in Figure 4A, and for commercially available fulvic acid, temperature-dependent surface tensions are shown in Figure 4B. It is important to note that the Simp-XGB, MACCS-XGB, and MF-XGB models were constrained during model training in such a way that σ_i° decreased monotonically with respect to T but not forcing a particular constant slope. σ_i° exhibited an approximately linear dependence on temperature across this range (consistent with typical expectations, e.g., eq 7). σ_i° decreased by approximately 18 mJ m^{-2} over the shown temperature range (200 to 350 K) in the case of the MF-XGB model for both compounds (i.e., a slope of ~ -0.12 mJ m^{-2} K^{-1}). For the MACCS-XGB model and the Simp-XGB model, σ_i° decreased by slightly more than 15 mJ m^{-2} for both compounds over the shown temperature range (200 to 350 K). In the case of humic acid, all three models predict slightly

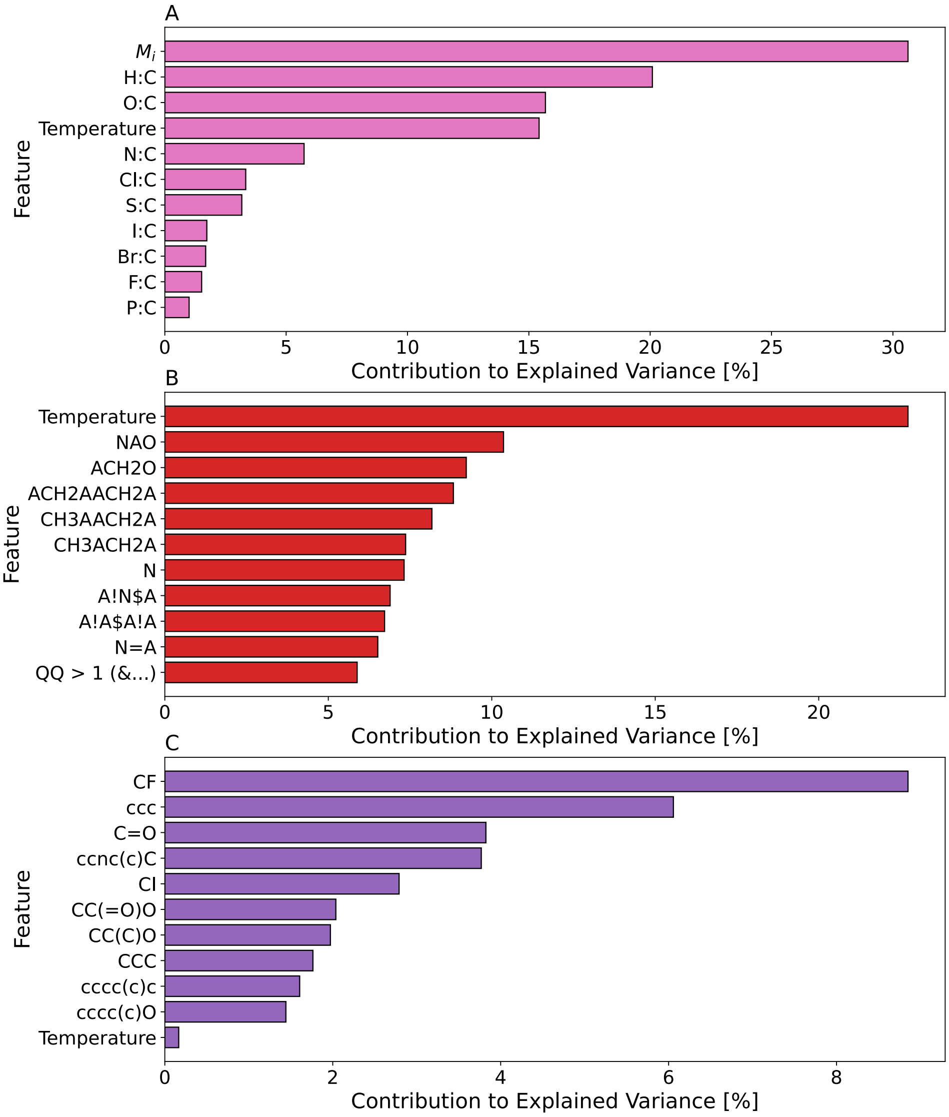


Figure 3. Percentage of variance in the XGB models explained by temperature and (A) simplified molecular properties, (B) the ten most important MACCS keys, and (C) the SMILES for the substructures corresponding to the ten most important Morgan fingerprint bit vectors. In panel (B), the symbol 'A' represents any element, '!' represents a nonring bond, '$' represents any bond in a ring system, '=' represents a double bond, and 'Q' represents any element that is neither carbon nor hydrogen.

distinct σ_i° values for the same compound, with the MF-XGB model yielding the highest value followed by the MACCS-XGB model and then the simplified model. For fulvic acid, the MACCS-XGB and Simp-XGB models predict similar values, while the MF-XGB model has lower predicted values for σ_i°. Also shown in Figure 4A are measured surface tensions of aqueous humic acid solutions at 295 K.[98,99] In Figure 4A, all three models have predicted surface tensions lower than those of the solution, with MF-XGB most closely replicating the experimental values. Figure 4B also shows two measured surface tensions of aqueous fulvic acid at 298 K.[98] For this case, MACCS-XGB slightly overpredicts $\sigma^\circ_{\text{fulvic acid}}$, with Simp-XGB best replicating the measurements and MF-XGB substantially underpredicting $\sigma^\circ_{\text{fulvic acid}}$ by nearly 10 mJ m^{-2}. Note that none of the plotted measurements represent pure humic or fulvic acid, respectively; the experimental data were taken from the highest concentrations reported in their respective works. For the aqueous humic acid solution, this concentration was 10.7 g L^{-1}[98] and 1.0 g L^{-1}.[99] For the aqueous fulvic acid solution, the concentration was 10.7 g L^{-1} as well.[99] Therefore, these values ought to be thought of as upper bounds for $\sigma^\circ_{\text{humic acid}}$ and $\sigma^\circ_{\text{fulvic acid}}$.

The 100 $\sigma^\circ - T$ data point pairs with the highest error for each of the Simp-XGB, MACCS-XGB, and MF-XGB models were extracted from the test data set to determine if there were additional shared features that may have contributed to their poor performance. Figure S3A shows the residuals of these compounds compared to reported values of σ_i°. We find that all three categories of inputs tend to overpredict σ_i° when the reported σ_i° value is lower than 20 mJ m^{-2}. Likewise the models tend to underpredict σ_i° when the reported value is between 20 and 50 mJ m^{-2}. Figure S3B shows model error as a function of T; in this case, there does not appear to be a systematic bias in the error of the poorest performing cases. Of these paired data, the mean temperatures were 332.9 K for the Simp-XGB model, 338.0 K for the MACCS-XGB model, and 330.3 K for the MF-XGB model. In the case of the simplified inputs, there was a statistically significant (Brunner-Munzel p =

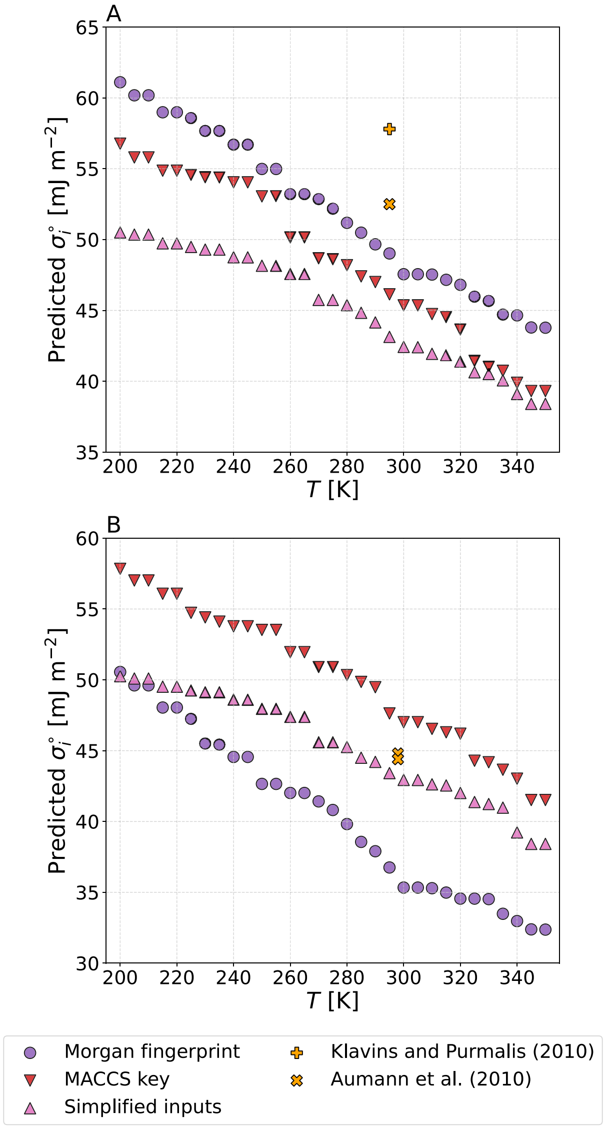


Figure 4. Temperature dependence of σ_i° for two biomass burning tracer species not found in the test data: (A) a representative humic acid structure (C1C2C=CC1C(C2C(=O)O)(C(=O)O)[N+](=O)[O−]) and (B) a representative fulvic acid structure (CC1(CC2=C(CO1)C(=O)C3=C(O2)C=C(C(=C3C(=O)O)O)O)O)O). The temperature range reflects temperatures that may occur within a smoke plume and the troposphere. We note that the reported measurement values were taken from aqueous solutions at the highest concentration of humic or fulvic acid, respectively.[98,99]

6×10^{-13} and Student's t test $p = 6 \times 10^{-4}$) decrease in the mean molar mass of the poorest performing predictions in comparison to the remainder of the test data. Likewise, there was also a weakly statistically significant decrease in the mean Br:C ($p = 0.075$) as described by the Brunner−Munzel test, although the Student's t test suggested a more significant relationship $p = 0.0017$. The Brunner−Munzel test also suggested that there was a statistically significant decrease in the I:C ($p = 2.94 \times 10^{-7}$) and P:C ($p = 2.45 \times 10^{-6}$) ratios from the rest of the test data. In other words, compounds containing I or P were more poorly predicted than compounds that did not contain these elements. No other elemental ratio exhibited a significant difference. The average T was slightly higher in the poorest performing data; however, it is important to note that the Student's t test statistical p-value for this was only weakly significant ($p = 0.06$). For a complete list of Student's t test and Brunner−Munzel test results for the inputs to the simplified model, please see Table S4. There were no significant differences between the mean T of the poorest performing data and the remainder of the test data in the MACCS-XGB ($p = 0.0977$) and MF-XGB models ($p = 0.883$).

For the more complex model inputs (MACCS keys and Morgan Fingerprints), similarities between poorly performing compounds were analyzed. Similarity scores between individual pairs of SMILES within the 100 poorest performing compounds were calculated using eq 24. The distance or dissimilarity between two compounds can be computed as $1 - S_{A,B}$. Compounds were then grouped hierarchically based on their distances from one another. Figures S4 and S5 show these hierarchical clusters for the MACCS-XGB and MF-XGB models. It can be seen that most compounds are more dissimilar than similar with only a small fraction of the compounds plotted clustered with distances below 0.5, marked by colored branches in those figures. It is also important to note that there are minimal differences between the clustering outputs.

As an example of the utility of the model developed in this work, the effect of using ML-generated surface tension values of organics in a mixture, instead of assuming a single constant value for all organics, is now explored. A system of 21 surrogates comprised of isoprene-derived multigeneration oxidation and/or fragmentation products was generated based on predictions by the Master Chemical Mechanism[94,95] and used as input with the models described in this work. For a complete list of the selected isoprene-derived oxidation and fragmentation products predicted by the MCM and their SMILES, see Table 2. The predicted pure-component surface tensions were used as inputs along with the AIOMFAC-based bulk−surface partitioning model of Schmedding and Zuend[14] to predict the critical supersaturation for mixed organic−inorganic aerosol particles with a water-free (dry) diameter of 50 nm. The particles also contain ammonium sulfate with the organic dry mass fraction of 0.73. The σ_i° of ammonium sulfate was predicted using the approach of Dutcher et al.[97] Figure 5A shows the Köhler curves using model-predicted σ_i° values for the isoprene-derived system. Figure 5B shows the Köhler curves using model-predicted σ_i° values for an analogous system comprising α-pinene ozonolysis products and ammonium sulfate with a water-free (dry) diameter of 50 nm and the organic dry mass fraction set to 0.8. For a complete list of the selected α-pinene ozonolysis products predicted by the MCM and their SMILES, see Table 3. These systems were analyzed to determine the computational cost of running each model from a given list of SMILES using a single core of an Intel i5−8265U CPU with 8GB of RAM. It took 0.018 s to generate the simplified model inputs from the SMILES of the 14 α-pinene products (Table 3) and to predict σ_i° using the Simp-XGB model. To generate MACCS keys and run the MACCS-XGB model from the same SMILES, it took 0.029 s. To compute Morgan fingerprints and run the MF-XGB model, it took 0.032 s. Much of this time was taken up by generating the input values from SMILES with RDKit v2024.3.5, as the time it took to run each model afterward was 0.0059 s for Simp-XGB, 0.0029 s for MACCS-XGB, and 0.0026 s for MF-XGB. It is important to note that the number of times any of the XGB models need to be called is variable depending on the user's intended application. For an AIOMFAC-based gas−particle

Table 2. Predicted Organic Surface Tensions for the Isoprene SOA Surrogate System as Shown in Figure 5A at T = 298 K[b]

MCM Name	SMILES	Morgan fingerprint σ_i° [mJ m^{-2}]	MACCS keys σ_i° [mJ m^{-2}]	Simplified inputs σ_i°[mJ m^{-2}]
IEB1OOH	OCC(O)C(C)(OO)C=O	22.90	39.92	44.69
IEB2OOH	OOC(C=O)C(C)(O)CO	25.24	38.33	44.69
C59OOH	OCC(=O)C(C)(CO)OO	45.07	36.68	44.69
IEC1OOH	OCC(=O)C(C)(CO)OO	45.07	36.68	44.69
C58OOH	O=CC(O)C(C)(CO)OO	36.78	33.90	44.69
IEPOXA	CC(O)(CO)C1CO1	28.22	36.40	35.22
C57OOH	OCC(O)C(C)(OO)C=O	22.90	39.92	44.69
IEPOXC	CC1(CO1)C(O)CO	35.52	36.40	35.22
HIEB1OOH	OCC(O)C(CO)(OO)C=O	37.78	45.31	44.71
INDOOH	OCC(ON(=O)=O)C(C)(CO)OO	43.45	39.85	44.90
IEACO3H	CC(O)(C1CO1)C(=O)OO	35.47	36.34	46.44
C525OOH	OCC(=O)C(CO)(CO)OO	54.89	44.48	44.71
HIEB2OOH	OOC(C=O)C(O)(CO)CO	34.14	45.79	44.71
IEC2OOH	OCC(=O)C(C)(OO)C=O	43.76	38.42	46.44
INAOOH	OCC(C)(OO)C(O)CON(=O)=O	40.73	39.93	44.90
C510OOH	O=CC(O)C(C)(OO)CON(=O)=O	36.43	38.89	50.10
INB1OOH	OCC(OO)C(C)(CO)ON(=O)=O	33.35	39.85	44.90
IECCO3H	CC1(CO1)C(O)C(=O)OO	33.82	35.94	46.44
INCOOH	OCC(OO)C(C)(O)CON(=O)=O	28.75	39.93	44.90
INB2OOH	OOCC(O)C(C)(CO)ON(=O)=O	37.22	39.97	44.90
Tetrol dimer	CC(O)(CO)C(O)COC(C)(CO)C(O)CO	23.20	37.48	41.55
$(NH_4)_2SO_4$	[NH4+].[NH4+].[O-]S(=O)(=O)[O-]		184.5[a]	

[a]Calculated using the approach of Dutcher et al.[97] [b]Isoprene SOA species and SMILES were taken from the Master Chemical Mechanism (MCM).[94,95]

partitioning calculation with coupled bulk−surface partitioning at constant temperature, a pure-component surface tension calculation need only be called one time at the start of the calculation.

Figures 5A and 5B also show a case where it is assumed that all organic species have the same surface tension as water and the case where all organic species have a surface tension of 35 mJ m^{-2}.[9,12] All inorganic electrolytes use the values reported by Dutcher et al.[97] in all systems. In the isoprene system, there is more variability between the predicted solution surface tensions shown in Figure 5C and thus more variation in the estimated critical supersaturations. In the case of the α-pinene system, there is less variability between the predicted solution surface tensions shown in Figure 5D, and all three models lead to similar predictions of the critical supersaturation for cloud droplet activation. It is important to note that the equilibrium bulk−surface partitioning framework laid out in Schmedding and Zuend[14] allows the predicted σ to vary in a range exceeding the highest and lowest σ_i° values, which may occur when the surface exhibits highly nonideal mixing, e.g. under dilute conditions. This is shown in Figure 5D where $\sigma > \sigma_w^\circ$ occurs near the point of cloud droplet activation. Furthermore, interactions between liquid−liquid phase separation, the three-dimensional configuration of a particle, such as whether it is of core−shell or partially engulfed morphology, and the selected treatment of interfacial tension may reduce this behavior in the bulk−surface partitioning framework.[14,19]

As an additional example of the model's ability to capture temperature-dependent effects, the same system as shown in Figures 5A and Figure 5C is calculated at 268 K. These results are shown in Figure S6. It is important to note that temperature also impacts the coupled gas−particle and bulk−surface partitioning of organic compounds, such as isoprene degradation products; therefore, changes in the cloud activation conditions (such as required supersaturation) in this case may not strictly be due to the σ_i° values of individual organic species.

DISCUSSION

Previous attempts to predict σ_i° have relied on empirical parametrizations, as mentioned in the Introduction. These equations all rely on various molecular properties, including but not limited to T_c, T_b, P_c, H_{vap}, ρ_l, ρ_v, $\mathcal{R}^\star$, which must be measured or predicted indirectly through application of additional models. This limits the utility of such surface tension models for real-world applications in environmental systems, wherein the pure-component properties of many compounds are poorly constrained due to the complex mechanisms by which they form and the wide variety of multifunctional organic compounds encountered in the atmosphere.[100] Additionally, many of these empirical models are trained only on specific compound classes. This likewise leads to limits on the ability of these models to predict σ_i° for compounds outside of the classes for which they are trained. These limitations are evident in Figure 2, wherein the three types of inputs discussed in this study (simplified inputs, MACCS key, and Morgan fingerprint) all yield more accurate predictions in comparison to the empirical parametrizations. It should be noted that although the models developed in this study were never trained on the test data, the empirical models that were used for comparison were never trained on these data either. It is therefore possible that these empirical models may show better performance when they are used with the compound classes and data sets for which they were trained.

The models discussed in this study rely on the structural properties of molecules along with the temperature to predict σ_i°. The most detailed description of molecular structure comes in the form of the Morgan fingerprints; however, such an approach has several drawbacks which limit its utility. The foremost among these is that the most accurate model, MF-

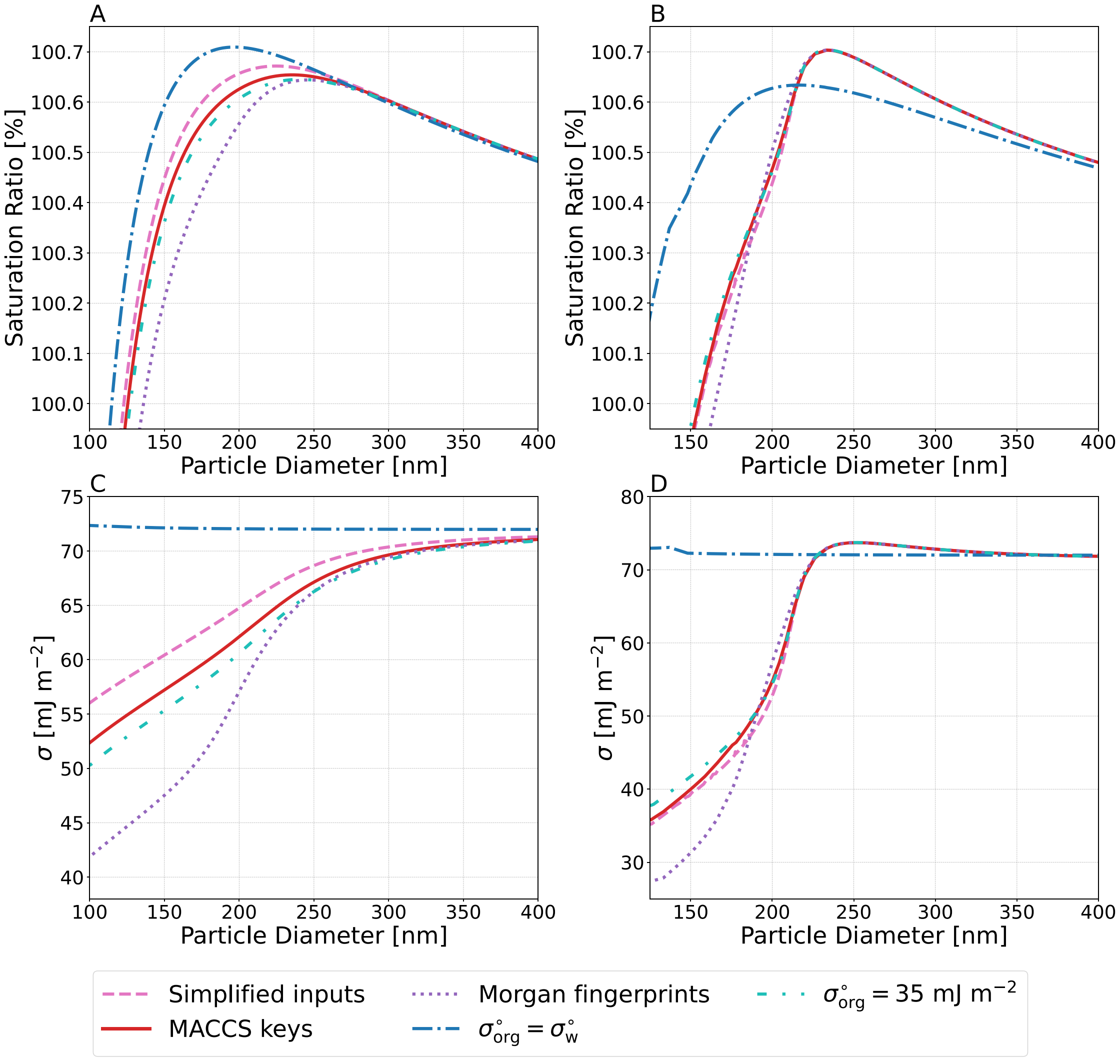


Figure 5. Predicted Köhler curves for (A) the isoprene oxidation products system and (B) the α-pinene oxidation products system at 298 K for particles with a dry diameter of 50 nm. The dry mass fraction of organic species is (A) 0.73 and (B) 0.8, with ammonium sulfate as the remainder. Also shown are Köhler curves generated with the assumption that all organic species have the same pure-component surface tension as water ($\sigma^\circ_{org} = \sigma^\circ_w$) or the assumption used in other studies that σ°_{org} = 35 mJ m^{-2}.[9,12] Panels (C) and (D) show the solution surface tension corresponding to panels (A) and (B) as a function of particle size.

XGB, performs only marginally better than the next most detailed series of inputs, MACCS-XGB, despite having over ten times the number of inputs for molecular descriptors (2048 vs 166). Of equal importance, MACCS keys are inherently human readable, whereas Morgan fingerprints rely on complex descriptions of the relative positions of each atom in a molecule simultaneously. This may make model analysis in terms of the importance of individual substructures mathematically difficult. Thus, it is recommended to use the MACCS-XGB model in the case where SMILES are known and the Simp-XGB model in cases where SMILES are unknown, but the elemental ratios and molar weight of a compound are readily available. It is also possible that other types of model inputs may perform well in predicting surface tension. Functional group-based approaches, such as one analogous to that used by AIOMFAC to describe organic compounds, are an additional option. It should be noted that such approaches may not capture the three-dimensional structure and relative orientations of functional groups as well as the MACCS keys or Morgan fingerprints. Likewise, some compounds with more reactive functional groups, such as peroxyacids, may be difficult to isolate in a laboratory setting. Thus, accurate surface tension measurements of these compounds are difficult to find.

CONCLUSIONS

In this work, three different approaches for characterizing organic molecules and predicting their temperature-dependent pure-component surface tensions were compared with one another. For each approach to molecular characterization, four different ML approaches were used: extreme gradient boosting, random forests, decision trees, and K-nearest neighbors. We find that the extreme gradient boosting approach results in the highest R^2 in comparison to observations, regardless of which method of molecular characterization for input is used. From each of the three approaches, the most important molecular features were extracted. The molecular properties in the data set used in this work with the highest importance were functional groups that contained nitrogen and oxygen along with ring structures. More broadly speaking, a combination of M_i, T, the O:C ratio, and the N:C ratio was able to explain

Table 3. Predicted Organic Surface Tensions for the α-Pinene SOA Surrogate System at T = 298 K[b]

MCM Name	SMILES	Morgan fingerprint σ_i° [mJ m^{-2}]	MACCS keys σ_i° [mJ m^{-2}]	Simplified inputs σ_i° [mJ m^{-2}]
C107OOH	O=CCC1CC(OO)(C(=O)C)C1(C)C	32.61	33.10	40.84
PINONIC	OC(=O)CC1CC(C(=O)C)C1(C)C	32.85	34.90	39.61
C97OOH	OCC1CC(OO)(C(=O)C)C1(C)C	33.33	34.65	40.24
C108OOH	O=CCC(CC(=O)C(=O)C)C(C)(C)OO	22.70	32.43	41.23
C89CO2H	O=CCC1CC(C(=O)O)C1(C)C	30.87	33.98	38.64
PINIC	OC(=O)CC1CC(C(=O)O)C1(C)C	31.22	36.45	39.99
C921OOH	OCC(=O)C1(OO)CC(CO)C1(C)C	49.38	36.66	39.22
C109OOH	OOCC(=O)C1CC(CC=O)C1(C)C	32.83	33.60	40.84
C812OOH	OCC1CC(OO)(C(=O)O)C1(C)C	37.90	36.96	44.96
HOPINONIC	OCC(=O)C1CC(CC(=O)O)C1(C)C	43.59	36.85	40.84
C811OH	OCC1CC(C(=O)O)C1(C)C	31.61	35.33	38.93
C813OOH	OCC(CC(=O)C(=O)O)C(C)(C)OO	30.85	37.20	45.37
ALDOL dimer	CC(=O)C(=O)CC(C(C=O)=CCC1CC(C(O)=O)C1(C)C)C(C)(C)OO	23.48	34.21	39.75
ESTER dimer	CC1(C)C(CC1C(O)=O)CC(=O)OCC(=O)C2CC(CC(O)=O)C2(C)C	31.26	34.09	39.75
$(NH_4)_2SO_4$	[NH4+].[NH4+].[O-]S(=O)(=O)[O-]		184.5[a]	

[a] Calculated using the approach of Dutcher et al.[97] [b] α-Pinene SOA species and SMILES codes were taken from the Master Chemical Mechanism (MCM).

80.1% of the variance in a simplified model which only relied on molecular ratios, M_i, and T as inputs. Such a model is useful in cases wherein the exact structure of an individual chemical species is not readily available. In the case where more detailed structural information is available in the form of SMILES, we find that using MACCS keys as inputs provides the best balance of model performance and input simplicity.

To demonstrate the importance of accurately characterizing temperature-dependent pure-component surface tension values, a Köhler curve was generated for an isoprene SOA surrogate system. It was found that the inclusion of the ML-based approaches for surface tension led to substantial changes in the predicted droplet surface tension evolution during hygroscopic growth compared to the frequently used implicit assumption of $\sigma_{\mathrm{org}}^\circ = \sigma_w^\circ$. The use of appropriate pure-component values is also shown to impact in the critical supersaturation necessary for cloud droplet activation. It is noted that some atmospherically relevant compounds, such as peroxyacids and peroxyacyl nitrates, are not well represented in the data set used in this work and that further measurements of the pure-component surface tensions of such compounds at atmospherically relevant temperatures may be necessary to improve model performance for these compound classes.

ASSOCIATED CONTENT

Data Availability Statement

The training and test data used in this work, the Simp-XGB, MF-XGB, and MACCS-XGB models, and the code used to optimize model hyperparameters and train models can be found at the following Zenodo archive: 10.5281/zenodo.13936980.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.4c00291.

> Additional information on the compound classes used to train the various models described throughout the work along with figures for the analysis of the poorest performing compounds from each type of molecular descriptor (PDF)

AUTHOR INFORMATION

Corresponding Author

Andreas Zuend − *Department of Atmospheric and Oceanic Sciences, McGill University, Montreal, Quebec H3A 0B9, Canada*; orcid.org/0000-0003-3101-8521; Email: andreas.zuend@mcgill.ca

Authors

Ryan Schmedding − *Department of Atmospheric and Oceanic Sciences, McGill University, Montreal, Quebec H3A 0B9, Canada*; orcid.org/0009-0003-0958-2676

Mees Franssen − *Department of Atmospheric and Oceanic Sciences, McGill University, Montreal, Quebec H3A 0B9, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.4c00291

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Rachel Wegener and Riley McCue for their assistance with the CryoCloud JupyterBook. This project was undertaken with financial support by the government of Canada through the federal Department of Environment and Climate Change (grant no. GCXE20S049) and the Natural Sciences and Engineering Research Council of Canada (NSERC grant no. RGPIN-2021-02688). RS was also supported by a doctoral scholarship from the Fonds de Recherche du QubecNature et technologies (FRQNT, scholarship no. 314186).

REFERENCES

(1) Aitken, J. XII.−On Dust, Fogs, and Clouds. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh* **1881**, *30*, 337−368.

(2) Twomey, S. Pollution and the planetary albedo. *Atmospheric Environment (1967)* **1974**, *8*, 1251−1256.

(3) Köhler, H. The nucleus in and the growth of hygroscopic droplets. *Trans. Faraday Soc.* **1936**, *32*, 1152−1161.

(4) Sorjamaa, R.; Svenningsson, B.; Raatikainen, T.; Henning, S.; Bilde, M.; Laaksonen, A. The role of surfactants in Köhler theory reconsidered. *Atmos. Chem. Phys.* **2004**, *4*, 2107−2117.

(5) Nozière, B.; Baduel, C.; Jaffrezo, J.-L. The dynamic surface tension of atmospheric aerosol surfactants reveals new aspects of cloud activation. *Nat. Commun.* **2014**, *5*, 3335.

(6) Gérard, V.; Nozière, B.; Baduel, C.; Fine, L.; Frossard, A. A.; Cohen, R. C. Cationic, and Nonionic Surfactants in Atmospheric Aerosols from the Baltic Coast at Askweden: Implications for Cloud Droplet Activation. *Environ. Sci. Technol.* **2016**, *50*, 2974−2982.

(7) Petters, S. S.; Petters, M. D. Surfactant effect on cloud condensation nuclei for two-component internally mixed aerosols. *Journal of Geophysical Research: Atmospheres* **2016**, *121*, 1878−1895.

(8) Ruehl, C. R.; Davies, J. F.; Wilson, K. R. An interfacial mechanism for cloud droplet formation on organic aerosols. *Science* **2016**, *351*, 1447−1450.

(9) Ovadnevaite, J.; Zuend, A.; Laaksonen, A.; Sanchez, K. J.; Roberts, G.; Ceburnis, D.; Decesari, S.; Rinaldi, M.; Hodas, N.; Facchini, M. C.; Seinfeld, J. H.; O'Dowd, C. Surface tension prevails over solute effect in organic-influenced cloud droplet activation. *Nature* **2017**, *546*, 637−641.

(10) Kroflič, A.; Frka, S.; Simmel, M.; Wex, H.; Grgić, I. Size-Resolved Surface-Active Substances of Atmospheric Aerosol: Reconsideration of the Impact on Cloud Droplet Formation. *Environ. Sci. Technol.* **2018**, *52*, 9179−9187.

(11) Malila, J.; Prisle, N. L. A Monolayer Partitioning Scheme for Droplets of Surfactant Solutions. *Journal of advances in modeling earth systems* **2018**, *10*, 3233−3251.

(12) Davies, J. F.; Zuend, A.; Wilson, K. R. Technical note: The role of evolving surface tension in the formation of cloud droplets. *Atmospheric Chemistry and Physics* **2019**, *19*, 2933−2946.

(13) Gérard, V.; Noziere, B.; Fine, L.; Ferronato, C.; Singh, D. K.; Frossard, A. A.; Cohen, R. C.; Asmi, E.; Lihavainen, H.; Kivekäs, N.; Aurela, M.; Brus, D.; Frka, S.; Cvitešić Kušan, A. Concentrations and Adsorption Isotherms for Amphiphilic Surfactants in PM1 Aerosols from Different Regions of Europe. *Environ. Sci. Technol.* **2019**, *53*, 12379−12388.

(14) Schmedding, R.; Zuend, A. A thermodynamic framework for bulk-surface partitioning in finite-volume mixed organic-inorganic aerosol particles and cloud droplets. *Atmospheric Chemistry and Physics* **2023**, *23*, 7741−7765.

(15) Sebastiani, F.; Campbell, R. A.; Rastogi, K.; Pfrang, C. Nighttime oxidation of surfactants at the air-water interface: effects of chain length, head group and saturation. *Atmospheric Chemistry and Physics* **2018**, *18*, 3249−3268.

(16) Pfrang, C.; Sebastiani, F.; Lucas, C. O. M.; King, M. D.; Hoare, I. D.; Chang, D.; Campbell, R. A. Ozonolysis of methyl oleate monolayers at the air-water interface: oxidation kinetics, reaction products and atmospheric implications. *Phys. Chem. Chem. Phys.* **2014**, *16*, 13220−13228.

(17) Knopf, D. A.; Forrester, S. M. Freezing of Water and Aqueous NaCl Droplets Coated by Organic Monolayers as a Function of Surfactant Properties and Water Activity. *J. Phys. Chem. A* **2011**, *115*, 5579−5591.

(18) Donaldson, D. J.; Vaida, V. The Influence of Organic Films at the Air-Aqueous Boundary on Atmospheric Processes. *Chem. Rev.* **2006**, *106*, 1445−1461.

(19) Schmedding, R.; Zuend, A. The role of interfacial tension in the size-dependent phase separation of atmospheric aerosol particles. *Atmospheric Chemistry and Physics* **2025**, *25*, 327−346.

(20) Johansson, J. H.; Salter, M. E.; Acosta Navarro, J. C.; Leck, C.; Nilsson, E. D.; Cousins, I. T. Global transport of perfluoroalkyl acids via sea spray aerosol. *Environ. Sci.: Processes Impacts* **2019**, *21*, 635−649.

(21) Cheng, M.; Li, Y.; Kuwata, M. Hysteresis in Water Content of Ultrafine Glassy Organic Aerosol Particles. *Journal of Geophysical Research: Atmospheres* **2024**, *129*, e2024JD041440.

(22) Wokosin, K. A.; Schell, E. L.; Faust, J. A. Emerging investigator series: surfactants, films, and coatings on atmospheric aerosol particles: a review. *Environ. Sci.: Atmos.* **2022**, *2*, 775−828.

(23) Topping, D. O.; McFiggans, G. B.; Kiss, G.; Varga, Z.; Facchini, M. C.; Decesari, S.; Mircea, M. Surface tensions of multi-component mixed inorganic/organic aqueous systems of atmospheric significance: measurements, model predictions and importance for cloud activation predictions. *Atmospheric Chemistry and Physics* **2007**, *7*, 2371−2398.

(24) Vepsäläinen, S.; Calderón, S. M.; Prisle, N. L. Comparison of six approaches to predicting droplet activation of surface active aerosol - Part 2: strong surfactants. *Atmos. Chem. Phys.* **2023**, *23*, 15149.

(25) Kleinheins, J.; Shardt, N.; El Haber, M.; Ferronato, C.; Nozière, B.; Peter, T.; Marcolli, C. Surface tension models for binary aqueous solutions: a review and intercomparison. *Phys. Chem. Chem. Phys.* **2023**, *25*, 11055−11074.

(26) Berry, M. V. The molecular mechanism of surface tension. *Physics Education* **1971**, *6*, 79.

(27) Zuend, A.; Marcolli, C.; Luo, B. P.; Peter, T. A thermodynamic model of mixed organic-inorganic aerosols to predict activity coefficients. *Atmospheric Chemistry and Physics* **2008**, *8*, 4559−4593.

(28) Zuend, A.; Marcolli, C.; Booth, A. M.; Lienhard, D. M.; Soonsin, V.; Krieger, U. K.; Topping, D. O.; McFiggans, G.; Peter, T.; Seinfeld, J. H. New and extended parameterization of the thermodynamic model AIOMFAC: calculation of activity coefficients for organic-inorganic mixtures containing carboxyl, hydroxyl, carbonyl, ether, ester, alkenyl, alkyl, and aromatic functional groups. *Atmospheric Chemistry and Physics* **2011**, *11*, 9155−9206.

(29) Kleinheins, J.; Marcolli, C.; Dutcher, C. S.; Shardt, N. A unified surface tension model for multi-component salt, organic, and surfactant solutions. *Phys. Chem. Chem. Phys.* **2024**, *26*, 17521−17538.

(30) Bzdek, B. R.; Power, R. M.; Simpson, S. H.; Reid, J. P.; Royall, C. P. Precise, contactless measurements of the surface tension of picolitre aerosol droplets. *Chem. Sci.* **2016**, *7*, 274−285.

(31) Bzdek, B. R.; Reid, J. P.; Malila, J.; Prisle, N. L. The surface tension of surfactant-containing, finite volume droplets. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 8335−8343.

(32) Bain, A.; Lalemi, L.; Croll Dawes, N.; Miles, R. E. H.; Prophet, A. M.; Wilson, K. R.; Bzdek, B. R. Surfactant Partitioning Dynamics in Freshly Generated Aerosol Droplets. *J. Am. Chem. Soc.* **2024**, *146*, 16028−16038.

(33) Bain, A.; Ghosh, K.; Prisle, N. L.; Bzdek, B. R. Surface-Area-to-Volume Ratio Determines Surface Tensions in Microscopic, Surfactant-Containing Droplets. *ACS Central Science* **2023**, *9*, 2076−2083.

(34) Bain, A.; Prisle, N. L.; Bzdek, B. R. Model-Measurement Comparisons for Surfactant-Containing Aerosol Droplets. *ACS Earth and Space Chemistry* **2024**, *8*, 2244−2255.

(35) Eberhart, J. G. The Surface Tension of Binary Liquid Mixtures1. *J. Phys. Chem.* **1966**, *70*, 1183−1186.

(36) Connors, K. A.; Wright, J. L. Dependence of surface tension on composition of binary aqueous-organic solutions. *Anal. Chem.* **1989**, *61*, 194−198.

(37) Shardt, N.; Elliott, J. A. W. Model for the Surface Tension of Dilute and Concentrated Binary Aqueous Mixtures as a Function of Composition and Temperature. *Langmuir* **2017**, *33*, 11077−11085.

(38) Shardt, N.; Wang, Y.; Jin, Z.; Elliott, J. A. Surface tension as a function of temperature and composition for a broad range of mixtures. *Chem. Eng. Sci.* **2021**, *230*, 116095.

(39) Sprow, F. B.; Prausnitz, J. M. Surface tensions of simple liquid mixtures. *Trans. Faraday Soc.* **1966**, *62*, 1105−1111.

(40) Antonow, G. N. Sur la tension superficielle à la limite de deux couches. *J. Chim. Phys.* **1907**, *5*, 372−385.

(41) Girifalco, L. A.; Good, R. J. A Theory for the Estimation of Surface and Interfacial Energies. I. Derivation and Application to Interfacial Tension. *J. Phys. Chem.* **1957**, *61*, 904−909.

(42) Hyvärinen, A.-P.; Lihavainen, H.; Gaman, A.; Vairila, L.; Ojala, H.; Kulmala, M.; Viisanen, Y. Surface Tensions and Densities of

Oxalic, Malonic, Succinic, Maleic, Malic, and cis-Pinonic Acids. *Journal of Chemical & Engineering Data* **2006**, *51*, 255−260.

(43) Riipinen, I.; Koponen, I. K.; Frank, G. P.; Hyvärinen, A.-P.; Vanhanen, J.; Lihavainen, H.; Lehtinen, K. E. J.; Bilde, M.; Kulmala, M. Adipic and Malonic Acid Aqueous Solutions: Surface Tensions and Saturation Vapor Pressures. *J. Phys. Chem. A* **2007**, *111*, 12995−13002.

(44) Lee, H. D.; Estillore, A. D.; Morris, H. S.; Ray, K. K.; Alejandro, A.; Grassian, V. H.; Tivanski, A. V. Direct Surface Tension Measurements of Individual Sub-Micrometer Particles Using Atomic Force Microscopy. *J. Phys. Chem. A* **2017**, *121*, 8296−8305.

(45) Macleod, D. B. On a relation between surface tension and density. *Trans. Faraday Soc.* **1923**, *19*, 38−41.

(46) Sugden, S. CXLI.−The influence of the orientation of surface molecules on the surface tension of pure liquids. *J. Chem. Soc., Trans.* **1924**, *125*, 1167−1177.

(47) Log, A. M.; Diky, V.; Huber, M. L. Assessment of a Parachor Model for the Surface Tension of Binary Mixtures. *Int. J. Thermophys.* **2023**, *44*, 110.

(48) Firoozabadi, A.; Katz, D. L.; Soroosh, H.; Sajjadian, V. A. Surface Tension of Reservoir Crude-Oil/Gas Systems Recognizing the Asphalt in the Heavy Fraction. *SPE Reservoir Engineering* **1988**, *3*, 265−272.

(49) Escobedo, J.; Mansoori, G. A. Surface-tension prediction for liquid mixtures. *AIChE J.* **1998**, *44*, 2324−2332.

(50) Weinaug, C. F.; Katz, D. L. Surface Tensions of Methane-Propane Mixtures. *Industrial & Engineering Chemistry* **1943**, *35*, 239−246.

(51) Hugill, J.; Van Welsenes, A. Surface tension: a simple correlation for natural gas + conensate systems. *Fluid Phase Equilib.* **1986**, *29*, 383−390.

(52) Quayle, O. R. The Parachors of Organic Compounds. An Interpretation and Catalogue. *Chem. Rev.* **1953**, *53*, 439−589.

(53) Escobedo, J.; Mansoori, G. A. Surface tension prediction for pure fluids. *AIChE J.* **1996**, *42*, 1425−1433.

(54) Lu, J.-F.; Fu, D.; Liu, J.-C.; Li, Y.-G. Application of density functional theory for predicting the surface tension of pure polar and associating fluids. *Fluid Phase Equilib.* **2002**, *194*−*197*, 755−769.

(55) Fu, D.; Lu, J.-F.; Liu, J.-C.; Li, Y.-G. Prediction of surface tension for pure non-polar fluids based on density functional theory. *Chem. Eng. Sci.* **2001**, *56*, 6989−6996. Festschrift in honor of Professor T.-M. Guo.

(56) Tang, X.; Gross, J. Density functional theory for calculating surface tensions with a simple renormalization formalism for the critical point. *J. Supercrit. Fluids* **2010**, *55*, 735−742.

(57) Guggenheim, E. A. The Principle of Corresponding States. *J. Chem. Phys.* **1945**, *13*, 253−261.

(58) Lielmezs, J.; Herrick, T. New surface tension correlation for liquids. *Chemical Engineering Journal* **1986**, *32*, 165−169.

(59) Brock, J. R.; Bird, R. B. Surface tension and the principle of corresponding states. *AIChE J.* **1955**, *1*, 174−177.

(60) Aleem, W.; Mellon, S. N.; Sufian, M.; Mutalib, I. A.; Subbarao, D. A Model for the Estimation of Surface Tension of Pure Hydrocarbon Liquids. *Petroleum Science and Technology* **2015**, *33*, 1908−1915.

(61) Gharagheizi, F.; Eslamimanesh, A.; Sattari, M.; Mohammadi, A. H.; Richon, D. Development of corresponding states model for estimation of the surface tension of chemical compounds. *AIChE J.* **2013**, *59*, 613−621.

(62) Sanjuán, E.; Parra, M.; Pizarro, M. Development of models for surface tension of alcohols through symbolic regression. *J. Mol. Liq.* **2020**, *298*, 111971.

(63) Randová, A.; Bartovská, L. Group contribution method: Surface tension of linear and branched alkanes. *Fluid Phase Equilib.* **2016**, *429*, 166−176.

(64) Soori, T.; Rassoulinejad-Mousavi, S. M.; Zhang, L.; Rokoni, A.; Sun, Y. A machine learning approach for estimating surface tension based on pendant drop images. *Fluid Phase Equilib.* **2021**, *538*, 113012.

(65) Rafie, S.; Hajipour, M.; Delijani, E. B. Modeling hydrocarbon surface tension using MLP and RBF neural networks and evolutionary optimization algorithms. *Petroleum Science and Technology* **2023**, *41*, 1622−1640.

(66) Ojaki, H. A.; Lashkarbolooki, M.; Movagharnejad, K. Checking the performance of feed-forward and cascade artificial neural networks for modeling the surface tension of binary hydrocarbon mixtures. *Journal of the Iranian Chemical Society* **2023**, *20*, 655−667.

(67) Mousavi, S.-P.; Atashrouz, S.; Nait Amar, M.; Hadavimoghaddam, F.; Mohammadi, M.-R.; Hemmati-Sarapardeh, A.; Mohaddespour, A. Modeling surface tension of ionic liquids by chemical structure-intelligence based models. *J. Mol. Liq.* **2021**, *342*, 116961.

(68) Lazzús, J. A.; Cuturrufo, F.; Pulgar-Villarroel, G.; Salfate, I.; Vega, P. Estimating the Temperature-Dependent Surface Tension of Ionic Liquids Using a Neural Network-Based Group Contribution Method. *Ind. Eng. Chem. Res.* **2017**, *56*, 6869−6886.

(69) Lashkarbolooki, M.; Bayat, M. Prediction of surface tension of liquid normal alkanes, 1-alkenes and cycloalkane using neural network. *Chem. Eng. Res. Des.* **2018**, *137*, 154−163.

(70) Pierantozzi, M.; Mulero, A.; Cachadina, I. Surface tension of liquid organic acids: An artificial neural network model. *Molecules* **2021**, *26*, 1636.

(71) Pazuki, G. R.; N, M.; Sahranavard, L. Prediction of Surface Tension of Pure Hydrocarbons by An Artificial Neural Network System. *Petroleum Science and Technology* **2011**, *29*, 2384−2396.

(72) Xu, T.; Khanghah, M. A.; Du, Y. Toward prediction of surface tension of branched n-alkanes using ANN technique. *Petroleum Science and Technology* **2019**, *37*, 127−134.

(73) Roosta, A.; Setoodeh, P.; Jahanmiri, A. Artificial Neural Network Modeling of Surface Tension for Pure Organic Compounds. *Ind. Eng. Chem. Res.* **2012**, *51*, 561−566.

(74) Yee, L. C.; Wei, Y. C. Current modeling methods used in QSAR/QSPR. *Statistical modelling of molecular descriptors in QSAR/QSPR* **2012**, *2*, 1−31.

(75) Uddin, S.; Lu, H. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *PLoS One* **2024**, *19*, e0301541.

(76) Chen, C.-H.; Tanaka, K.; Kotera, M.; Funatsu, K. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics* **2020**, *12*, 19.

(77) Boldini, D.; Grisoni, F.; Kuhn, D.; Friedrich, L.; Sieber, S. A. Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics* **2023**, *15*, 73.

(78) Cicciarelli, B. A.; Hatton, T. A.; Smith, K. A. Dynamic Surface Tension Behavior in a Photoresponsive Surfactant System. *Langmuir* **2007**, *23*, 4753−4764.

(79) Consonni, V.; Ballabio, D.; Todeschini, R. In *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development*; Roy, K., Ed.; Academic Press, 2023; pp 303−327.

(80) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(81) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5*, 107−113.

(82) Orsi, M.; Reymond, J.-L. One chiral fingerprint to find them all. *Journal of Cheminformatics* **2024**, *16*, 53.

(83) Jasper, J. J. The Surface Tension of Pure Liquid Compounds. *J. Phys. Chem. Ref. Data* **1972**, *1*, 841−1010.

(84) OpenEye Scientific Software, I. *OEChem Toolkit*, Version 2.3.0. 2024; https://pubchem.ncbi.nlm.nih.gov.

(85) Nti, I. K.; Nyarko-Boateng, O.; Aning, J.; et al. Performance of machine learning algorithms with different K values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science* **2021**, *13*, 61−71.

(86) Victoria, A. H.; Maragatham, G. Automatic tuning of hyperparameters using Bayesian optimization. *Evolving Systems* **2021**, *12*, 217−223.

(87) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery, New York, NY, USA, 2019.

(88) Snow, T.; Millstein, J.; Scheick, J.; Sauthoff, W.; Leong, W. J.; Colliander, J.; Pérez, F.; Munroe, J.; Felikson, D.; Sutterley, T.; Siegfried, M. *CryoCloud JupyterBook*. version 2023.01.06. 2023.

(89) Landrum, G. et al. *RDKit*. version 2024.03.5. 2024.

(90) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, 2016.

(91) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.

(92) Yaws, C. L. *Yaws' thermophysical properties of chemicals and hydrocarbons*, [electronic ed.] ed.; Knovel: Norwich, N.Y., 2009.

(93) Fligner, M. A; V, J. S.; Blower, P. E. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.

(94) Jenkin, M. E.; Young, J. C.; Rickard, A. R. The MCM v3.3.1 degradation scheme for isoprene. *Atmospheric Chemistry and Physics* **2015**, *15*, 11433−11459.

(95) Rastak, N.; et al. Microphysical explanation of the RH-dependent water affinity of biogenic organic aerosol and its importance for climate. *Geophys. Res. Lett.* **2017**, *44*, 5167−5177.

(96) Jenkin, M. E.; Wyche, K. P.; Evans, C. J.; Carr, T.; Monks, P. S.; Alfarra, M. R.; Barley, M. H.; McFiggans, G. B.; Young, J. C.; Rickard, A. R. Development and chamber evaluation of the MCM v3.2 degradation scheme for β-caryophyllene. *Atmospheric Chemistry and Physics* **2012**, *12*, 5275−5308.

(97) Dutcher, C. S.; Wexler, A. S.; Clegg, S. L. Surface Tensions of Inorganic Multicomponent Aqueous Electrolyte Solutions and Melts. *J. Phys. Chem. A* **2010**, *114*, 12216−12230.

(98) Aumann, E.; Hildemann, L.; Tabazadeh, A. Measuring and modeling the composition and temperature-dependence of surface tension for organic solutions. *Atmos. Environ.* **2010**, *44*, 329−337.

(99) Klavins, M.; Purmalis, O. Humic substances as surfactants. *Environmental Chemistry Letters* **2010**, *8*, 349−354.

(100) Goldstein, A. H.; Galbally, I. E. Known and unexplored organic constituents in the earth's atmosphere. *Environ. Sci. Technol.* **2007**, *41*, 1514−1521.