Author(s): Hyttinen, Noora; Li, Linjie; Hallquist, Mattias; Wu, Cheng

Title: Machine Learning Model to Predict Saturation Vapor Pressures of Atmospheric Aerosol Constituents

Year: 2024

Version: Published version

Please cite the original version:

Hyttinen, N., Li, L., Hallquist, M., & Wu, C. (2024). Machine Learning Model to Predict Saturation Vapor Pressures of Atmospheric Aerosol Constituents. ACS - ES & T Air, Early online. https://doi.org/10.1021/acsestair.4c00113

ACS ES&T Air

Open Access

pubs.acs.org/estair

Article

Machine Learning Model to Predict Saturation Vapor Pressures of Atmospheric Aerosol Constituents

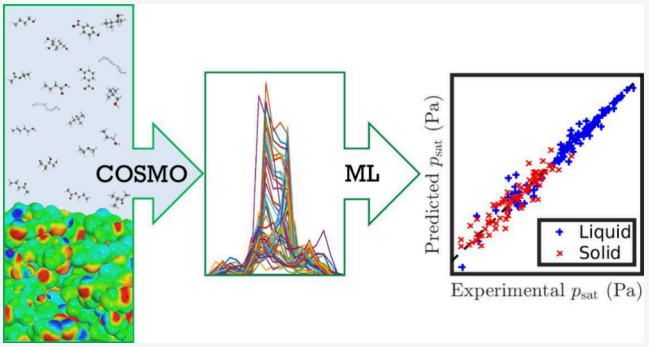
Published as part of ACS ES&T Air virtual special issue "Elevating Atmospheric Chemistry Measurements and Modeling with Artificial Intelligence".

Noora Hyttinen,* Linjie Li, Mattias Hallquist, and Cheng Wu

Cite This: https://doi.org/10.1021/acsestair.4c00113

Read Online

ABSTRACT: We present a novel machine learning (ML) model for predicting saturation vapor pressures (p_{sat}), a physical property of use to describe transport, distribution, mass transfer, and fate of environmental toxins and contaminants. The ML model uses σ-profiles from the conductor-like screening model (COSMO) as molecular descriptors. The main advantages in using σ-profiles instead of other types of molecular representations are the relatively small size of the descriptor and the fact that the addition of new elements does not affect the size of the descriptor. The ML model was trained separately for liquid and solid compounds using experimental vapor pressures at various temperatures. The 95% confidence intervals of the error in the liquid- and solid-phase $\log_{10}(p_{sat}/\mathrm{Pa})$ are 1.02 and 1.4, respectively. Especially our solid-phase model outperforms all group-contribution models in predicting experimental sublimation pressures of solid compounds. To demonstrate its applicability, the model was used to predict p_{sat} of atmospherically relevant species, and the values were compared with those obtained from a new experimental method. Here, our model provided a tool for a better description of this critical property and gave a higher confidence in the measurements.



KEYWORDS: *COSMO, extreme minimal learning machine, σ-profile, liquid, solid, volatility*

■ INTRODUCTION

Saturation vapor pressure of organic compounds is a useful thermodynamic property in many applications. For example, saturation vapor pressures are needed to model the transport, distribution, mass transfer, and fate of environmental toxins and contaminants. In atmospheric research, saturation vapor pressure is used to model the gas-to-particle partitioning of organic compounds formed in the gas phase in order to estimate the growth rates of aerosol particles. Various measurement techniques exist for determining saturation vapor pressures. However, the determination of saturation vapor pressures of low volatility compounds is difficult, and often different measurements give orders of magnitude different results.[1] Additionally, it is not feasible to measure the saturation vapor pressures of all environmental contaminants and atmospheric trace gases.

Many empirical models exist for the estimation of saturation vapor pressures, varying from simple equations that require only knowledge on the elemental composition[2,3] to group-contribution models[4−7] that also consider various functional groups. More complex and time consuming quantum chemistry-based models, such as the conductor-like screening model for real solvents (COSMO-RS[8−10]), can even take the stereoisomer into account. Recently, the COSMO-RS model has been used to calculate saturation vapor pressures of complex organic compounds.[1,11−14] Currently, the most advanced parametrization of the COSMO-RS model is implemented in the commercially available BIOVIA COSMO*therm* program.[15]

Several studies have investigated the effect of conformer sampling on COSMO*therm* calculations of saturation vapor pressure.[1,12,14,16] Generally, different conformers can lead to orders of magnitude differences in saturation vapor pressure estimates. For example, Kurtén et al.[16] recommended selecting conformers of multifunctional compounds based on their intramolecular hydrogen bonding. Especially in the condensed phase, conformers that are able to interact with the

Received: May 27, 2024
Revised: July 11, 2024
Accepted: July 11, 2024

surrounding system are energetically more favorable.[17] Conversely, Stahn et al.[14] recommended selecting a set of lowest energy conformers for the COSMO*therm* calculations. Li et al.[1] found that conformers in the COSMO*base* (a cosmo-file database of common compounds) produced accurate saturation vapor pressure estimates of mono- and dicarboxylic acids and sugar alcohols. On the other hand, using the default cosmo-file generation procedure of the COSMO*conf* program[18] did not lead to adequate agreement between COSMO*therm*-estimated and experimental saturation vapor pressures of polyethylene glycols (PEG).[1]

These different findings highlight that the parametrization of COSMO*therm* is still biased toward the relatively simple compounds that have been used in the parametrization of the model. A more reliable and systematic way of including new compounds is therefore needed to be able to truly predict saturation vapor pressures of new compounds. For this reason, a systematic way to select optimal conformers for both the parametrization of the model and the prediction of new compounds is crucial.

With the development of machine learning (ML) techniques, ML models are quickly replacing quantum chemistry calculations and traditional thermodynamic models. Warnau et al. noted that empirical machine learning methods are currently outperforming COSMO*therm* in partition coefficient calculations.[19] Some studies have used conductor-like screening model (COSMO[20])-based ML techniques to predict various thermodynamic properties.[21−26] In these models, the quantum chemistry output from the COSMO model (i.e., σ-profile) is used to create a molecular representation of each compound for the ML model.

The aim of this study was to create a COSMO-based ML model (COSMO-ML) to predict the saturation vapor pressures of environmentally relevant multifunctional organic compounds. The created model was trained separately with experimental data for vapor pressures above both the liquid and solid phase to get a better estimate on the effect of the solid-to-liquid phase transition. Additionally, the study includes an evaluation of the model and a demonstration of its applicability to atmospherically relevant compounds, showing its usefulness in the description of the fate and behavior of low volatility organic environmental toxins.

METHODS

The ideal partial vapor pressure of compound i (p_i) can be calculated from the free energy of vaporization (ΔG_{vap})

$$p_i(T) = a_i(T)e^{-\Delta G_{\mathrm{vap},i}(T)/RT} \tag{1}$$

Here, R is the gas constant, T is the temperature, and a_i is the activity of the condensed-phase compound i in the mixture. For ideal pure condensed-phase compounds, $a = 1$ when the pure compound is selected as the reference state. For a real gas containing monomers, dimers, trimers, and even larger clusters, the presence of clusters in the gas phase should be considered in the calculation of the free energy of vaporization.

In the COSMO-RS model, the free energy of vaporization of a pure compound is derived from density functional theory (DFT) calculations. In practice, the free energies are calculated separately for the gas and condensed phases ($G^{(g)}(T)$ and $G^{(c)}(T)$, respectively). This leads to the following equation for the saturation vapor pressure (p_{sat}):

$$p_{\mathrm{sat}} = e^{[G^{(c)}(T)-G^{(g)}(T)]/RT} \tag{2}$$

The gas-phase free energy is obtained with a vacuum DFT calculation. The condensed-phase liquid free energy ($G^{(l)}(T)$) is calculated from the DFT COSMO energy (E_{COSMO}) and the chemical potential of the pure liquid compound ($\mu^{(l)}(T)$ from the COSMO-RS model).

$$G^{(l)}(T) = E_{\mathrm{COSMO}} + \mu^{(l)}(T) \tag{3}$$

On the other hand, the condensed-phase solid free energy ($G^{(s)}(T)$) is calculated from E_{COSMO}, the liquid-phase $\mu^{(l)}(T)$, and the free energy of fusion ($\Delta G_{\mathrm{fus}}(T)$).

$$G^{(s)}(T) = E_{\mathrm{COSMO}} + \mu^{(l)}(T) + \Delta G_{\mathrm{fus}}(T) \tag{4}$$

It is not possible to estimate $\Delta G_{\mathrm{fus}}(T)$ using computational methods. Instead, $\Delta G_{\mathrm{fus}}(T)$ must be derived experimentally. For simplicity, we will use $\mu(T)$ to describe $\mu^{(l)}(T)$ for the liquid phase and $\mu^{(l)}(T) + \Delta G_{\mathrm{fus}}(T)$ for the solid phase.

In this study, we use COSMO-ML instead of COSMO-RS to predict the values of $\mu(T)$. The target value predicted with the ML model will be calculated using $p_{\mathrm{sat}}(T)$ from experiments, and E_{COSMO} and $G^{(g)}$ from DFT calculations:

$$\mu(T) = RT \ln p_{\mathrm{sat}}(T) - E_{\mathrm{COSMO}} + G^{(g)} \tag{5}$$

Note that E_{COSMO} and $G^{(g)}$ are not temperature dependent, which means that the temperature dependence of p_{sat} is included in the $\mu(T)$ term. As μ is defined differently for liquid and solid compounds, the COSMO-ML model needs to be trained separately for liquid and solid compounds. It should also be noted that the use of experimental p_{sat} as training data means that predicted μ corresponds to an effective free energy of vaporization instead of the actual thermodynamic free energy of vaporization. Additionally, the COSMO-ML model will be trained using experimental saturation vapor pressures measured for real gas-phase mixtures containing monomers, dimers, trimers, etc., instead of ideal gas phases containing only monomers. The predicted values will therefore also take into account any effects of clustering in the gas-phase on the saturation vapor pressure.

The advantage of predicting μ, instead of p_{sat} directly, is that the range of possible values of μ is more narrow than that of p_{sat}. Similar approaches are often used to predict energies of molecules or molecular cluster of different sizes. For example, the energy is first computed at a low level of theory (fast calculation) and a machine learning method is used to predict the energy difference between the low level of theory and a desired high level of theory (slow calculation). Additionally, if the training data include compounds with an adequate range of μ values, even predictions outside the p_{sat} range of the training data can be accurate if the μ is within the range of the training data.

Machine Learning Model. The machine learning model used to predict chemical potentials is extreme minimal learning machine (EMLM[27]). The EMLM model is a kernel-based method that uses Euclidean distances as a similarity measure. From the training data ($\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times n_x}$, N data points in total, n_x features in the descriptor), K_{ref} reference points are selected to train the model. The reference points are selected based on the Euclidean distances of the descriptors, so that the reference points are spread evenly in the Euclidean distance space. Once the reference points have been selected, the

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX

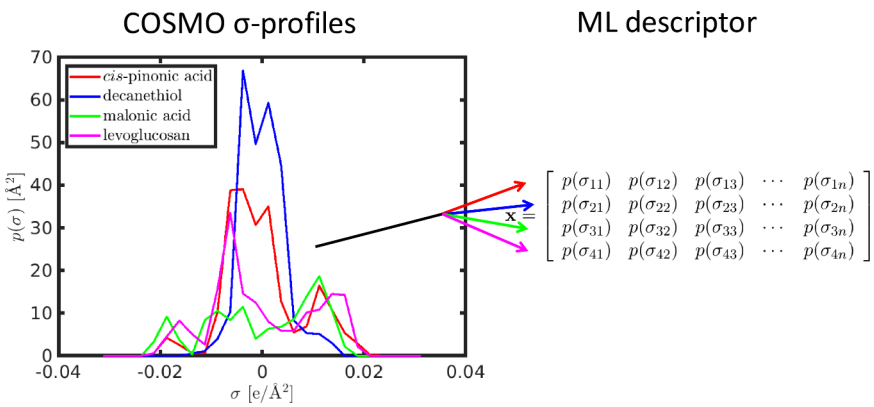


Figure 1. Molecular descriptors from COSMO σ-profiles ($p(\sigma)$).

Euclidean distances between all reference points are collected into a matrix $\mathbf{D} \in \mathbb{R}^{K_{\text{ref}} \times K_{\text{ref}}}$. To train the model, a regularized least-squares optimization problem will be solved to find the optimal $\mathbf{W}$ with

$$\min_{\mathbf{W} \in \mathbb{R}^{K_{\text{ref}} \times n_y}} J(\mathbf{W}) = \frac{1}{2N} \sum_{i=1}^{N} |\mathbf{d}_i^T \mathbf{W} - \mathbf{y}_i^T|^2 + \frac{\beta}{2K_{\text{ref}}} \sum_{i=1}^{K_{\text{ref}}} \sum_{j=1}^{n_y} |W_{ij}|^2 \tag{6}$$

Here, $\mathbf{d}_i \in \mathbb{R}^{K_{\text{ref}}}$ contains Euclidean distances between the ith input data point and all reference points; $\mathbf{W} \in \mathbb{R}^{K_{\text{ref}} \times n_y}$ is a weight matrix of the EMLM model; $\mathbf{y}_i$ is the experimental thermodynamic property value of data point i; β is a regularization parameter; and n_y is the number of target values for each data point (here $n_y = 1$).

The main advantage of using a distance based ML model, such as EMLM, is that overfitting is rarely an issue.[27,28] Additionally, the computational cost of EMLM is significantly lower than, for example, of neural networks or other deep learning methods. Another advantage of EMLM is that, unlike other kernel-based methods, EMLM only has 2 easy to optimize hyperparameters: (i) the number of reference points (K_{ref}) and (ii) the regularization parameter β describing how closely the model should be fitted to the training data target values. Since there are significant uncertainties in experimental thermodynamic property values, the value of β will be tested carefully. If the β value is too low, the model tries to represent all training data points perfectly, leading to over-fitting. On the other hand, higher β values allow larger differences between the training data and the model, accounting for the uncertainties of our experimental training data.

Descriptor. The input of the ML model was derived from the COSMO model. In the COSMO model, each conformer of a molecule is represented by a σ-surface. The σ-surface is a representation of the screening charge densities of a particular conformer of a molecule. The screening charge is the opposite of the surface charge. For the descriptor, we created σ-profiles (see an example in Figure 1) of each conformer. For the σ-profiles, the screening charge densities (charges/areas) of each compound were divided into bins. Each value of the σ-profile ($p(\sigma)$) is the total surface area (in Å^2) of the molecule that has the screening charge density corresponding to the bin. The bin size was optimized to find the best correlation between the predicted and experimental saturation vapor pressures. In addition to the σ-profile, the temperature was added as a single value to each of the descriptor vectors. If the vapor pressure of a compound was measured at multiple temperatures, all of the experimental data points were added for the compound with only the temperature changing in the descriptor. For the machine learning model, all features in the descriptor and the target value μ were scaled between −1 and 1. This ensures that features with larger absolute values in the descriptor are not prioritized in the training of the ML model.

The TURBOMOLE program package[29] and the BIOVIA COSMO*conf* program[18] were used to obtain the σ-profiles. First, a set of conformers was generated using the Spartan20 program[30] with systematic search algorithm and MMFF force field.[31] Generally, the shape of the carbon skeleton has only a small effect on the σ-profile (and chemical potential) of a conformer, because carbon chains mainly contribute to the σ-profile around the 0.00 e Å^{-2} charge density (see the σ-profile of decanethiol in Figure 1). On the contrary, the charge density of polar functional groups (i.e., negative and positive σ values for hydrogen bond donors and acceptors, respectively) depends strongly on the existence of intramolecular hydrogen bonds. The maximum number of conformers was therefore kept below 1000 by selecting only some of the torsions of long carbon chains for conformer sampling. The geometries of all conformers were optimized first at the BP/SV(P) and then at the BP/TZVP level of theory using TURBOMOLE. Duplicate conformers were omitted after both optimizations based on similarities in the geometries using the CLUSTER_GEO-CHECK algorithm of COSMO*conf*. The final single-point cosmo-files were calculated at the BP/def2-TZVPD-FINE//BP/def-TZVP level of theory. Gas-phase geometries were obtained by optimizing the BP/def-TZVP COSMO geometries in a vacuum, also at the BP/def-TZVP level of theory. The final gas-phase energies were calculated at the BP/def2-TZVPD level of theory in a vacuum.

In COSMO*therm*, the highest weight in the conformer distribution is given to the lowest free energy conformer. Here, a single conformer was selected to represent each compound in the COSMO-ML model. We used two different methods for selecting the condensed-phase conformer and two different options for gas-phase energy. For the condensed-phase conformer, we selected the one with the lowest free energy (calculated from the COSMO energy and the pure compound chemical potential at 298 K). Additionally, we tested omitting conformers with relative chemical potentials above 8 kJ/mol (about 2 kcal/mol) from the lowest chemical potential before selecting the lowest free energy conformer for the COSMO-ML calculation. This was done to avoid including conformers with high chemical potentials, as the COSMO model is known to overestimate the effect of intramolecular hydrogen bonds on the COSMO energy.[16] Generally, strong intramolecular

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX

hydrogen bonds (e.g., concerted hydrogen bonding of dicarboxylic acids) in the condensed phase increase the chemical potential of the pure compound. On the other hand, the COSMO energies of conformers containing intramolecular hydrogen bonds are significantly lower than those of conformers that contain no intramolecular hydrogen bonds. In COSMO-RS calculations, the gas-phase energy can be taken from the COSMO calculation (single-point energy calculation for the condensed-phase conformer), or from a separate gas-phase geometry optimization. Here, we tested using both the lowest gas-phase energy conformer (given in the energy-file, an output file of a vacuum calculation) and the single-point gas-phase energy of the selected COSMO conformer (given in the cosmo-file, an output file of a COSMO calculation). The comparison of the COSMO-ML model performance using the different input selection is shown in Table S1 of the Supporting Information. There is only a small difference between the COSMO conformer selection methods because most of the conformers are the same using both methods. Larger differences are seen between the two gas-phase energies. The best overall fit was found using the low relative chemical potential conformers, and the single-point gas-phase energy calculated for the condensed-phase geometry.

Training Data. Experimental saturation vapor pressure values were collected from published experimental studies. Details of the dataset used are given in Section S1, and Tables S2 and S3 of the Supporting Information. The training dataset contains equilibrium vapor pressures of 181 liquid-phase compounds and 112 solid-phase compounds, providing two models, i.e., one model for each phase. Some of the compounds are common for both phases. Sixty four of the liquid-phase compounds and 33 of the solid-phase compounds have experimental saturation vapor pressures measured at multiple temperatures. These compounds are used as test compounds only in model optimization. For testing of the final model, these compounds were used only as training data. In total, the training data contain 950 points of liquid-phase p_{sat} and 351 of solid-phase p_{sat}.

The training data have μ values (calculated using eq 5) ranging from −9.7 to 52.7 kJ/mol for the liquid compounds and from −33.5 to 41.3 kJ/mol for the solid compounds. This range of μ values is similar to COSMO*therm*-derived pure compound chemical potentials calculated for a large set of multifunctional oxidation products of α-pinene.[32] Most atmospherically relevant oxidized compounds are therefore within the training data of our model with regard to their μ values. The distribution of the training set in the μ-vs-T space is shown in Figure S1 of the Supporting Information.

■ RESULTS AND DISCUSSION

Model Optimization. In the model optimization, we tested the bin size of the descriptor σ-profile, the hyper-parameters of the EMLM model (β and K_{ref}), and the conformer selection. The parameters of the COSMO-ML model were optimized using the solid-phase data, which have fewer data points than the liquid-phase data. It is therefore more critical to find optimal model parameters in the solid-phase model than in the liquid-phase model. For cross-validation, we used the leave-one-out cross-validation, which is a k-fold cross-validation methods that uses the total number of data points as k ($k = N$), i.e., the model was trained N times by omitting a single data point at a time and predicting the p_{sat} of the omitted compound and temperature (here N = 334, some of the training data were added after the parameter optimization). This cross-validation maximizes the amount of training data during testing of the model. It should be noted that for the compounds with measured p_{sat} at multiple temperatures, this type of cross-validation will only show the ability of the COSMO-ML model to predict the effect of the temperature, as the σ-profile of the compound is included in the training of the model. The absolute error values in the test calculations of this section are therefore significantly lower than those of the final COSMO-ML model new compounds.

σ-Profile. In a σ-profile, the surface segments are divided into bins based on their charge density. The optimal bin size may vary, depending on the level of theory used for the COSMO calculations. With larger bin sizes, the results become more robust as small differences in σ values do not affect the assignment of the charge density in the descriptor. On the other hand, increasing the bin size will lead to loss of important differences between similar compounds.

We tested the effect of the bin size on the prediction ability of the COSMO-ML model. The bin size was varied from 0.001 to 0.006 e Å^{-2}, where the common side of the two central bins was always at the origo. The 95 % confidence interval of the prediction error as a function of the bin size is shown in Figure S2 of the Supporting Information. There is no large difference in the error with bin sizes between 0.0015 and 0.0055 e Å^{-2}. The lowest error is found at 0.0025 e Å^{-2}, with 26 bins in total (between −0.0325 and 0.0325 e Å^{-2}). A bin size of 0.0025 e Å^{-2} is therefore used in further calculations.

Regularization Parameter β. The 95 % confidence interval of the prediction error decreases when the value of β is decreased. However, there is no significant improvement below β = 0.01. In order to avoid over-fitting, β = 0.01 was selected for the final model.

Number of Reference Points K_{ref}. The reference points are selected based on the Euclidean distances between all data points so that the points are divided evenly in the Euclidean distance space. Figure S3 of the Supporting Information shows the convergence of the model with increasing K_{ref}. The model has converged with 50 % reference points, which corresponds to about 1 data point for each compound and an additional 2 data points for all compounds that have multiple temperature points in the data set. The main improvement achieved by the increasing of reference points above 50 % is in the prediction of temperature dependence of p_{sat} (black markers in Figure S3 of the Supporting Information). For our purpose of predicting vapor pressures of new compounds at atmospherically relevant temperatures, K_{ref} = 50 % (33 % for the liquid-phase model) is preferred to avoid bias toward compounds that are represented by multiple temperature points during the training of the model. However, for investigations of temperature dependence of vapor pressures, we recommend using K_{ref} = 100 %.

COSMO-ML Prediction Accuracy. The final testing of the COSMO-ML models was done the same way as the optimization of the model parameters. However, only compounds with one temperature point were included in the test data. Additionally, the p_{sat} values of all test compounds were measured around room temperature (295−303.4 K). We are therefore only testing the prediction ability of the COSMO-ML models for new compounds around room temperature. The COSMO-ML prediction error as a function of experimental (a) liquid-phase saturation vapor pressures and (b) solid-phase sublimation vapor pressures is shown in Figure 2.

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX

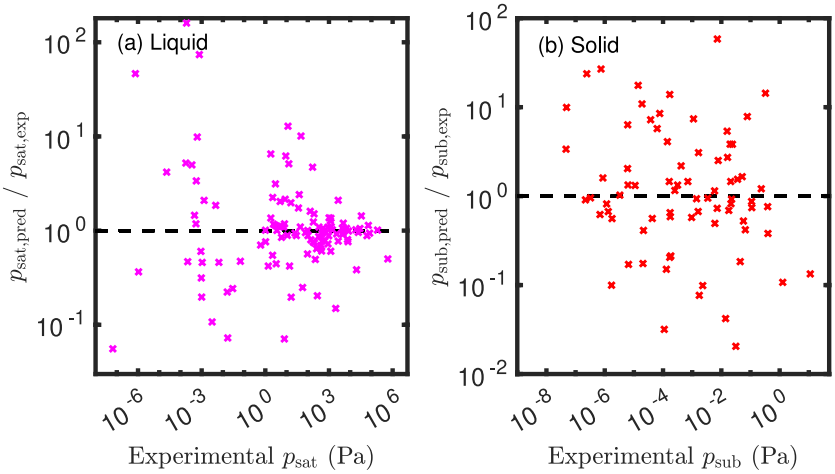


Figure 2. Prediction errors of the (a) liquid-phase and (b) solid-phase COSMO-ML models.

The 95% confidence interval of the prediction error of the liquid-phase p_{sat} is 1.02 orders of magnitude. The prediction accuracy is higher (0.67 orders of magnitude) in the high p_{sat} range (experimental p_{sat} > 0.1 Pa) than in the low p_{sat} range (1.8 orders of magnitude). The 95% confidence interval for the solid-phase p_{sat} prediction (mostly low p_{sat} range) is 1.4 orders of magnitude. Compounds that have low vapor pressures are generally relatively large and can exist in many conformers. Using a single conformer instead of an ensemble of favorable conformers to represent each compound may not be sufficient for complex multifunctional compounds, leading to larger errors in the model predictions. It should be noted that even small inaccuracies in the prediction of μ lead to large inaccuracies in p_{sat} due to the exponential relation. The 95 % confidence interval for the prediction error of μ (or free energy of vaporization) is only 4.1 and 7.7 kJ/mol (0.97 and 1.83 kcal/mol) for the liquid and solid models, respectively.

Some of the uncertainty in the low p_{sat} range predictions may be caused by errors in the experimental determination of saturation vapor pressures. For example, Wania et al.[33] commented on the measured p_{sat} of dinitronaphthalene and several dihydroxynaphthalenes, suggesting that their real liquid-phase p_{sat} are likely significantly higher than reported by Bannan et al.[34] None of the compounds measured by Bannan et al.[34] were included in the training data of our liquid-phase model. Our predicted liquid-phase p_{sat} are similar to those estimated using version 14 of COSMO*therm* (see Table S4 of the Supporting Information).[33] In addition to the disagreement in experimental and predicted p_{sat} of the dinitronaphthalene and dihydroxynaphthalenes, our model underestimates the liquid-phase p_{sat} of *para*-nitroaniline reported by Bannan et al.[34] by almost 3 orders of magnitude. On the other hand, our prediction agrees with the COSMO*therm*-estimated value.[33] The dinitronaphthalene, dihydroxynaphthalenes, and *para*-nitroaniline were therefore omitted from the training data of the solid-phase model as well. Including these aromatic compounds in the training data worsens the overall performance of the COSMO-ML model because the speculated error in the experimental p_{sat} is 2 orders of magnitude or higher.

Many experimental methods require high temperatures to measure the evaporation of low p_{sat} compounds. A large fraction of the training data is therefore high temperature p_{sat} (see Figure S1 of the Supporting Information). However, there are some compounds with relatively low and high calculated μ, which have measured p_{sat} values only at room temperature (295 or 298.15 K). For example, all of the high μ compounds in Figure S1 are polyethylene glycols (PEG). Here, PEG9 has a higher μ than any other solid-phase compound in our dataset (other PEGs are liquids). This leads to a large degree of extrapolation, and our model underestimating the p_{sat} of PEG9 by 5 orders of magnitude. PEG9 was therefore left out of the test set. In future calculations, the accuracy of the COSMO-ML model can be evaluated based on whether the predicted μ is within the limits of the corresponding training data in Figure S1.

The p_{sat} of a liquid phase is always higher than that of the corresponding solid phase at the same temperature below the melting point. This is due to the additional free energy needed for the solid-to-liquid phase transition. The trained liquid-phase COSMO-ML model was tested for the solid test compounds to investigate if the liquid-phase model predicts a higher p_{sat} than the solid-phase model. Most of the solid test compounds (around 92%) have predicted liquid-phase p_{sat} higher than the predicted solid-phase p_{sub}. This indicates that the independently trained models are able to find differences in the p_{sat} and p_{sub} of the same compound. For compounds that are within the size range of the training data of both the liquid- and solid-phase models (i.e., 6−28 nonhydrogen atoms), the effective free energy of fusion below the melting temperature can be estimated as the difference between μ of the liquid- and solid-phase predictions.

Comparison with Group-Contribution Models. Next, we compared our COSMO-ML predictions with existing group-contribution methods calculated using the UManSysProp code.[35] The models include SIMPOL,[5] EVAPORATION,[6] Myrdal and Yalkowsky,[7] and Nannoolal.[4] All of these group-contribution models are commonly used in atmospheric research to estimate the saturation vapor pressures of multifunctional organic compounds. For this comparison, we omitted test compounds that contain phosphorus and bromine, as these elements are not included in the group-contribution models. For simplicity, the p_{sat} values were calculated at 298.15 K for all compounds, even though the experimental temperatures vary between 295.0 and 303.4 K.

Figure S4 of the Supporting Information shows the performance of the models in estimating saturation vapor pressures of both solid and liquid compounds. Our COSMO-

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX

Table 1. COSMO-ML-Predicted, Experimental, and COSMO*therm*-Estimated Saturation Vapor Pressures (p_{sat} in Pa) of Δ^3-Carene Oxidation Products at 293.15 K

compound	molecular formula	COSMO-ML, solid	COSMO-ML, liquid	COSMO*therm*, liquid	experiment, Li et al.[36]
caric acid	$C_9H_{14}O_4$	1.3×10^{-5}	5.7×10^{-5}	2.2×10^{-4}	1.6×10^{-4}
caronic acid	$C_{10}H_{16}O_3$	1.2×10^{-4}	1.3×10^{-3}	6.9×10^{-3}	1.9×10^{-4}
OH-caronic acid	$C_{10}H_{16}O_4$	1.3×10^{-5}	2.6×10^{-4}	3.7×10^{-4}	2.6×10^{-4}

ML model outperforms all of the group-contribution models, especially for the solid compounds. Unlike our COSMO-ML model, none of the group-contribution models have been parametrized separately for solid compounds.

The 95% confidence interval in the errors of the group-contribution models vary from 1.22 (Nannoolal vapor pressures with Nannoolal boiling points) to 3.30 orders of magnitude (Nannoolal vapor pressures with Jopack−Reid boiling points) for the liquid compounds and from 3.04 (Myrdal−Yalkowsky vapor pressures with Joback−Reid boiling points) to 5.85 orders of magnitude (Myrdal−Yalkowsky vapor pressures with Nannoolal boiling points) for the solid compounds. The percentage of compounds that either under- or overestimate the experimental liquid-phase saturation vapor pressure value by more than 1 order of magnitude is 11−22% for the group-contribution models and 6% for our COSMO-ML model. The corresponding percentages for sublimation pressure are 47−88% and 15% for the group-contribution models and COSMO-ML, respectively. Figure S5 of the Supporting Information shows the fraction of test compounds as a function of the difference between experimental and calculated p_{sat} for each of the tested models.

Application to Atmospheric SOA Constituents. A potential application of our model is to estimate the vapor pressures of compounds that are hard to derive experimentally using pure authentic standards. As an example, Li et al.[36] recently estimated saturation vapor pressures of caric, caronic and OH-caronic acid from gas-to-particle partitioning coefficients using a filter inlet for gases and aerosols (FIGAERO) combined with a time-of-flight chemical ionization mass spectrometer (ToF-CIMS). These data were extracted from an oxidation experiment of Δ^3-carene. Our new model could then directly be applied as a comparison to these measurements. Here, we predicted saturation vapor pressures for the compounds proposed by Li et al.[36] using both the liquid- and solid-phase COSMO-ML models. Table 1 shows the COSMO-ML-derived and experimental p_{sat} values at 293.15 K. We additionally calculated the liquid-phase p_{sat} using the COSMO*therm* program with the newest BP_TZVPD_FINE_21 parametrization. In the COSMO*therm* calculation, we used sets of up to 10 lowest free energy conformers selected the same way as the conformers for the COSMO-ML models. For the gas phase, we used an equal number of the lowest gas-phase energy conformers.

Both COSMO-ML models agree with the experimentally determined p_{sat} values within about 1 order of magnitude. This is better than SIMPOL and EVAPORATION estimates.[36] The COSMO*therm*-estimated p_{sat} of caric and OH-caronic acid agree well with the experiments, while COSMO*therm* overestimates the experimental p_{sat} of caronic acid by a factor of 36. One major advantage of our COSMO-ML model compared to that of COSMO*therm* is that our solid-phase model can predict saturation vapor pressures of solid-phase compounds. In COSMO*therm*, additional experimental input is needed to derive the free energy of fusion of each compound. The experimental p_{sat} of caric and OH-caronic acid agree better with our liquid-phase COSMO-ML model, while the experimental p_{sat} of caronic acid agrees better with the solid-phase model. However, with the uncertainties of both the experiments and the models, it is not possible to determine the phase of the different acids measured by Li et al.[36] The solid-to-liquid p_{sat} ratio of the COSMO-ML models is between 5 and 10, which agrees with the observations of Booth et al.[37] for similar carboxylic acids. Thus, this application of the model illustrates its use in combination with new experimental methods to add a higher confidence in the estimation of thermodynamic data and our understanding of these properties.

Future Model Improvements. We have presented a working COSMO-ML model for predicting the saturation and sublimation vapor pressures of organic compounds. The cosmo-files and machine learning codes are included in https://doi.org/https://doi.org/10.23729/f2b13fc5-b3d1-49b4-a895-53e994a8218a for further development and use of the model. For future work, we recommend three main developments that may improve the accuracy of the model the most significantly.

A single conformer cannot be used to represent a realistic conformer distribution of a compound in the condensed-phase. Multifunctional compounds especially can have many different hydrogen bonding patterns, which affect the energies of the compound in the condensed and gas phases. Ideally, each compound would be represented by the weighted sum of the σ-profiles of a set of relevant conformers. Similar approach is used in the COSMO*therm* program.[15] The selection of appropriate conformer distributions in the COSMO-ML models will be further investigated in future work.

When using the COSMO-ML models, the size of the compound should be considered in order to not extrapolate outside the trained model. Currently, the training data consist of molecules with 3−30 (liquid) or 6−28 (solid) nonhydrogen atoms (mainly C, O, H, N, and S). However, the number of molecules with >20 nonhydrogen atoms is only 5 and 6 in the training data of the liquid- and solid-phase models, respectively. Larger compounds may have σ-profile ($p(\sigma)$) values that are outside the ranges of the training data compounds. The size range of compounds and the accuracy of the model can be further improved in the future by adding more compounds to the training data set. Additionally, the prediction accuracy of the temperature dependence of p_{sat} can be improved with more extensive training data. Accurate temperature dependence predictions will enable the estimation of the heat and entropy of sublimation and vaporization.

We have demonstrated the applicability of the COSMO σ-profiles in predicting saturation vapor pressures of multifunctional organic compounds. Here we have used a single machine learning model for our predictions. In the future, other machine learning techniques, such as neural networks, may be tested.

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Data Availability Statement

The cosmo-files used to train and test the model, and codes for predicting saturation vapor pressures can be accessed at Hyttinen, N. Machine Learning Model to Predict Saturation Vapor Pressures of Atmospheric Aerosol Constituents; University of Jyväskylä, 2024. DOI: 10.23729/f2b13fc5-b3d1-49b4-a895-53e994a8218a.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsestair.4c00113.

Table S1: Conformer selection; Section S1, Tables S2 and S3: Training data; Figure S1: Distribution of the training data; Figure S2: Bin size; Figure S3: Reference point convergence; Figures S4 and S5: Comparison with group-contribution models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Noora Hyttinen − *Department of Chemistry, Nanoscience Center, University of Jyväskylä, FI-40014 Jyväskylä, Finland; Department of Chemistry and Molecular Biology, University of Gothenburg, SE-40530 Gothenburg, Sweden;* Present Address: Finnish Meteorological Institute, FI-70211 Kuopio, Finland; orcid.org/0000-0002-6025-5959; Email: noora.hyttinen@fmi.fi

Authors

Linjie Li − *Department of Chemistry and Molecular Biology, University of Gothenburg, SE-40530 Gothenburg, Sweden;* orcid.org/0000-0003-0508-4947

Mattias Hallquist − *Department of Chemistry and Molecular Biology, University of Gothenburg, SE-40530 Gothenburg, Sweden;* orcid.org/0000-0001-5691-1231

Cheng Wu − *Department of Chemistry and Molecular Biology, University of Gothenburg, SE-40530 Gothenburg, Sweden;* orcid.org/0000-0001-9648-2865

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestair.4c00113

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the financial contribution from the Research Council of Finland, Grant No. 338171, and CSC-IT Center for Science, Finland, for computational resources. We thank the Swedish Research Council VR (2023-04520). It is a contribution to the Swedish strategic research area ModElling the Regional and Global Earth system, MERGE. We thank Dr. Silvia Calderón for helpful discussions.

■ REFERENCES

(1) Li, Z.; Hyttinen, N.; Vainikka, M.; Tikkasalo, O.-P.; Schobesberger, S.; Yli-Juuti, T. Saturation vapor pressure characterization of low-volatility organic compounds using isothermal aerosol evaporation. *Atmos. Chem. Phys.* **2023**, *23*, 6863−6877.

(2) Donahue, N. M.; Epstein, S. A.; Pandis, S. N.; Robinson, A. L. A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics. *Atmos. Chem. Phys.* **2011**, *11*, 3303−3318.

(3) Peräkylä, O.; Riva, M.; Heikkinen, L.; Quéléver, L.; Roldin, P.; Ehn, M. Experimental investigation into the volatilities of highly oxygenated organic molecules (HOMs). *Atmos. Chem. Phys.* **2020**, *20*, 649−669.

(4) Nannoolal, Y.; Rarey, J.; Ramjugernath, D. Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions. *Fluid Phase Equilib.* **2008**, *269*, 117−133.

(5) Pankow, J. F.; Asher, W. E. SIMPOL. 1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. *Atmos. Chem. Phys.* **2008**, *8*, 2773−2796.

(6) Compernolle, S.; Ceulemans, K.; Müller, J.-F. EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions. *Atmos. Chem. Phys.* **2011**, *11*, 9431−9450.

(7) Myrdal, P. B.; Yalkowsky, S. H. Estimating pure component vapor pressures of complex organic molecules. *Ind. Eng. Chem. Res.* **1997**, *36*, 2494−2499.

(8) Klamt, A. Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. *J. Phys. Chem.* **1995**, *99*, 2224−2235.

(9) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074−5085.

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

(11) Hyttinen, N.; Elm, J.; Malila, J.; Calderón, S. M.; Prisle, N. L. Thermodynamic properties of isoprene- and monoterpene-derived organosulfates estimated with COSMO*therm*. *Atmos. Chem. Phys.* **2020**, *20*, 5679−5696.

(12) Hyttinen, N.; Wolf, M.; Rissanen, M. P.; Ehn, M.; Peräkylä, O.; Kurtén, T.; Prisle, N. L. Gas-to-Particle Partitioning of Cyclohexene- and α-Pinene-Derived Highly Oxygenated Dimers Evaluated Using COSMO*therm*. *J. Phys. Chem. A* **2021**, *125*, 3726−3738.

(13) Roldin, P.; Ehn, M.; Kurtén, T.; Olenius, T.; Rissanen, M. P.; Sarnela, N.; Elm, J.; Rantala, P.; Hao, L.; Hyttinen, N.; et al. The role of highly oxygenated organic molecules in the Boreal aerosol-cloud-climate system. *Nature Comm.* **2019**, *10*, 1−15.

(14) Stahn, M.; Grimme, S.; Salthammer, T.; Hohm, U.; Palm, W.-U. Quantum chemical calculation of the vapor pressure of volatile and semi volatile organic compounds. *Environ. Sci.: Processes Impacts* **2022**, *24*, 2153−2166.

(15) *BIOVIA COSMOtherm*, Release 2021; Dassault Systèmes, 2021, http://www.3ds.com

(16) Kurtén, T.; Hyttinen, N.; D'Ambro, E. L.; Thornton, J.; Prisle, N. L. Estimating the saturation vapor pressures of isoprene oxidation products $C_5H_{12}O_6$ and $C_5H_{10}O_6$ using COSMO-RS. *Atmos. Chem. Phys.* **2018**, *18*, 17589−17600.

(17) Hyttinen, N.; Prisle, N. L. Improving solubility and activity estimates of multifunctional atmospheric organics by selecting conformers in COSMO*therm*. *J. Phys. Chem. A* **2020**, *124*, 4801−4812.

(18) *BIOVIA COSMOconf, 2021*; Dassault Systèmes, 2021, http://www.3ds.com.

(19) Warnau, J.; Wichmann, K.; Reinisch, J. COSMO-RS predictions of logP in the SAMPL7 blind challenge. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 813−818.

(20) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc. Perkin Trans.* **1993**, *II*, 799−805.

(21) Zhao, Y.; Huang, Y.; Zhang, X.; Zhang, S. A quantitative prediction of the viscosity of ionic liquids using $S_{\sigma-\text{profile}}$ molecular descriptors. *Phys. Chem. Chem. Phys.* **2015**, *17*, 3761−3767.

(22) Járvás, G.; Kontos, J.; Babics, G.; Dallos, A. A novel method for the surface tension estimation of ionic liquids based on COSMO-RS theory. *Fluid Phase Equilib* **2018**, *468*, 9−17.

(23) Kang, X.; Liu, X.; Li, J.; Zhao, Y.; Zhang, H. Heat capacity prediction of ionic liquids based on quantum chemistry descriptors. *Ind. Eng. Chem. Res.* **2018**, *57*, 16989−16994.

(24) Díaz, I.; Rodríguez, M.; González-Miquel, M.; González, E. J. COSMO-derived descriptors applied in ionic liquids physical property modelling using machine learning algorithms. *Comput. Aided Chem. Eng.* **2018**, *43*, 121−126.
(25) Kang, X.; Liu, C.; Zeng, S.; Zhao, Z.; Qian, J.; Zhao, Y. Prediction of Henry's law constant of CO_2 in ionic liquids based on S_{EP} and $S_{\sigma-profile}$ molecular descriptors. *J. Mol. Liq.* **2018**, *262*, 139−147.
(26) Nordness, O.; Kelkar, P.; Lyu, Y.; Baldea, M.; Stadtherr, M. A.; Brennecke, J. F. Predicting thermophysical properties of dialkylimidazolium ionic liquids from sigma profiles. *J. Mol. Liq.* **2021**, *334*, 116019.
(27) Kärkkäinen, T. Extreme Minimal Learning Machine: Ridge Regression with Distance-Based Basis. *Neurocomputing* **2019**, *342*, 33−48.
(28) Hämäläinen, J.; Alencar, A. S. C.; Kärkkäinen, T.; Mattos, C. L. C.; Júnior, A. H. S.; Gomes, J. P. P. Minimal learning machine: Theoretical results and clustering-based reference point selection. *J. Mach. Learn. Res.* **2020**, *21*, 1−29.
(29) *TURBOMOLE*, Version 7.7; a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, TURBOMOLE GmbH: Karlsruhe, 2010.
(30) *Spartan'20*; Wavefunction Inc.: Irvine, CA, 2020.
(31) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490−519.
(32) Hyttinen, N.; Pihlajamäki, A.; Häkkinen, H. Machine Learning for Predicting Chemical Potentials of Multifunctional Organic Compounds in Atmospherically Relevant Solutions. *J. Phys. Chem. Lett.* **2022**, *13*, 9928−9933.
(33) Wania, F.; Awonaike, B.; Goss, K.-U. Comment on "Measured Saturation Vapor Pressures of Phenolic and Nitro-Aromatic Compounds. *Environ. Sci. Technol.* **2017**, *51*, 7742−7743.
(34) Bannan, T. J.; Booth, A. M.; Jones, B. T.; O'Meara, S.; Barley, M. H.; Riipinen, I.; Percival, C. J.; Topping, D. Measured Saturation Vapor Pressures of Phenolic and Nitro-aromatic Compounds. *Environ. Sci. Technol.* **2017**, *51*, 3922−3928.
(35) Topping, D.; Barley, M.; Bane, M. K.; Higham, N.; Aumont, B.; Dingle, N.; McFiggans, G. UManSysProp v1. 0: an online and open-source facility for molecular property prediction and atmospheric aerosol calculations. *Geosci. Model Dev.* **2016**, *9*, 899−914.
(36) Li, L.; Thomsen, D.; Wu, C.; Priestley, M.; Iversen, E. M.; Tygesen Skønager, J.; Luo, Y.; Ehn, M.; Roldin, P.; Pedersen, H. B.; Bilde, M.; Glasius, M.; Hallquist, M. Gas-to-Particle Partitioning of Products from Ozonolysis of Δ^3-Carene and the Effect of Temperature and Relative Humidity. *J. Phys. Chem. A* **2024**, *128*, 918−928.
(37) Booth, A. M.; Montague, W. J.; Barley, M. H.; Topping, D. O.; McFiggans, G.; Garforth, A.; Percival, C. J. Solid state and sub-cooled liquid vapour pressures of cyclic aliphatic dicarboxylic acids. *Atmos. Chem. Phys.* **2011**, *11*, 655−665.

https://doi.org/10.1021/acsestair.4c00113
ACS EST Air XXXX, XXX, XXX−XXX