atmosphere

MDPI

Article

Machine Learning Approach for the Estimation of Henry's Law Constant Based on Molecular Descriptors

Atta Ullah [1], Muhammad Shaheryar [2] and Ho-Jin Lim [1,3,*]

[1] School of Architecture, Civil, Environmental and Energy Engineering, Kyungpook National University, Daegu 41566, Republic of Korea; 2018327287@knu.ac.kr
[2] School of Computer Science and Engineering, Kyungpook National University, Daegu 41566, Republic of Korea; shaheryar@knu.ac.kr
[3] Department of Environmental Engineering, Kyungpook National University, Daegu 41566, Republic of Korea
* Correspondence: hjlim@knu.ac.kr; Tel.: +82-53-950-7546; Fax: +82-53-550-6579

Abstract: In atmospheric chemistry, the Henry's law constant (HLC) is crucial for understanding the distribution of organic compounds across gas, particle, and aqueous phases. Quantitative structure–property relationship (QSPR) models described in scientific research are generally tailored to specific groups or categories of substances and are often developed using a limited set of experimental data. This study developed a machine learning model using an extensive dataset of experimental HLCs for approximately 1100 organic compounds. Molecular descriptors calculated using alvaDesc software (v 2.0) were used to train the models. A hybrid approach was adopted for feature selection, ensuring alignment with the domain knowledge. Based on the root mean squared error (RMSE) of the training and test data after cross-validation, Gradient Boosting (GB) was selected as a model for predicting HLC. The hyperparameters of the selected model were optimized using the automated hyperparameter optimization framework Optuna. The impact of features on the target variable was assessed using the SHapley Additive exPlanations (SHAP). The optimized model demonstrated strong performance across the training, evaluation, and test datasets, achieving coefficients of determination (R^2) of 0.96, 0.78, and 0.74, respectively. The developed model was used to estimate the HLC of compounds associated with carbon capture and storage (CCS) emissions and secondary organic aerosols.

Keywords: Henry's law constant; molecular descriptors; machine learning; atmospheric chemistry

check for updates

Citation: Ullah, A.; Shaheryar, M.; Lim, H.-J. Machine Learning Approach for the Estimation of Henry's Law Constant Based on Molecular Descriptors. *Atmosphere* **2024**, *15*, 706. https://doi.org/10.3390/atmos15060706

Academic Editor: Stephan Havemann

Received: 18 April 2024
Revised: 7 June 2024
Accepted: 11 June 2024
Published: 13 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The distribution of compounds among the gas, particle, and aqueous phases in the atmosphere is influenced by partition coefficients such as the Henry's law constant (HLC) [1]. Henry's law constant is defined as c/p under equilibrium conditions, where c represents the concentration of the species in the aqueous phase, and p is the partial pressure of the species in the gas phase. Compounds with high solubility can efficiently partition into the atmospheric aqueous phase, such as cloud water, from the gas phase [2]. In a field study conducted in China, Xuan et al. (2020) investigated the influence of HLC on the formation of oxidation products, specifically the partitioning of H_2O_2 between the atmospheric gas and liquid phases [3]. In cloud processing, where secondary organic aerosols are incorporated into cloud droplets, the partitioning of secondary organic aerosols and organic compounds between atmospheric particles, gas, and aqueous phases is characterized by HLC [4]. Experimentally determining HLC values is limited by factors such as solute adsorption in experimental systems, detection limits of analytical systems, potential interferences from impurities, and the challenges associated with measuring low concentrations [5–7]. Consequently, reliable estimation and prediction methods for determining HLC are critical.

Estimation methods, such as the ratio of vapor pressure to aqueous solubility and the quantity structure–property relationship (QSPR), are typically used to estimate HLC.

These methods provide estimates of species HLCs, which can be useful in assessing their contribution to environmental pollution [8]. Meylan and Howard (1991) used the bond contribution method to estimate the HLCs of 345 organic compounds. The model performed well in estimating the HLCs of hydrocarbons and monofunctional compounds (e.g., alkanes, alkenes, alcohols, aliphatic amines, etc.), achieving a RMSE and R^2 value of 0.52 and 0.94, respectively [9]. The method became inaccurate as the chemical structures became more complex. To address the proximity effects within molecules, Lin and Sandler (2002) developed a new hybrid quantum mechanics–group contribution method, which accounted for the contributions of different functional groups located at various positions on the same molecule [10]. This method achieved a RMSE of 0.34 for a dataset of 395 organic compounds. Gharagheizi et al. (2010) used a hybrid method that combined group contribution and feed-forward neural networks to estimate the HLC for 1940 pure organic compounds. The HLC values in the dataset were obtained by estimating methods rather than being solely based on experimental data. The RMSE and R^2 of the model were 0.1 and 0.9981, respectively, but the model also showed a standard deviation of more than 30% for a set of 14 compounds in the dataset [11]. In a recent study, Duchowicz et al. (2020) used a QSPR approach to predict HLC using a dataset of 530 diverse compounds, including pesticides, solvents, aromatic hydrocarbons, and persistent pollutants [12]. The developed model achieved a RMSE of 1.3 and R^2 of 0.76. Although the model's predictive performance was shown to be similar to that of Henry's Law Constant Program for Microsoft Windows (HENRYWIN), it is noteworthy that HENRYWIN's group contribution and bond contribution techniques may not be equally accurate for complex molecules.

Machine learning (ML) is an advanced and powerful tool for predicting the physical properties of chemical compounds. For instance, Vo Thanh et al. (2024) employed machine learning to predict hydrogen solubility in aqueous systems [13]. Additionally, Hou et al. (2022) developed a deep learning model based on Bidirectional long short-term memory with Channel and Spatial Attention network (BCSA) using the molecular Simplified Molecular Input Line Entry System (SMILES) to predict aqueous solubility [14]. Furthermore, machine learning aids in characterizing the absorption and adsorption kinetics of CO_2 in ionic liquids and metal–organic frameworks by predicting its HLC using Random Forest (RF), Multiple Linear Regression (MLP), and Support Vector Machine (SVM) [15–17]. Wang et al. (2020) used an Adaptive Neuro-Fuzzy Inference System (ANFIS) and a Least Squares Support Vector Machine (LSSVM) to predict HLC in water based on the molecular structure of compounds. The authors reported the mean squared error (MSE) and RMSE for the ANFIS model as 0.0072 and 0.0097, respectively. However, they did not provide details about the structural effect on HLC prediction [18].

Recent advancements in machine learning have significantly advanced the prediction of physical properties such as the HLC of organic compounds. However, challenges remain in accurately predicting HLC because of limited experimental data for model training and prediction limitations for complex molecules. While previous studies on HLC prediction have focused on developing QSPR models, they have often been constrained by a lack of experimental data, reliance on functional groups and bonds in molecules for predictions, and limited application in the prediction of HLC of intricate molecules. Our ML model addresses the shortcomings of previous studies by using reliable and experimentally measured HLCs for a broad range of compounds and extends its applicability to predict the HLCs of complex compounds related to carbon capture and storage (CCS) and atmospheric chemistry. Additionally, the model interprets the complex underlying chemical relationship between the target variable and input features, providing a broader understanding of the concept. In this study, we developed a ML model to predict HLC based on molecular descriptors. In addition to linear models such as Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net, we employed various boosting algorithms, including tree-based methods like eXtreme Gradient Boosting (XGB), Category Boosting (CatBoost), Light Gradient Boosted Machine (LightGBM), and Gradient Boosting (GB), as well as non-tree-based methods like Adaptive Boosting (AdaBoost). We used the embedded technique

for feature selection. The best-performing algorithm was selected and further optimized using Bayesian optimization. Subsequently, we applied the optimized model to the test data for validation. The model's performance was compared to that of the vapor pressure–aqueous solubility model. Finally, we applied the final model to predict the HLC values of different classes of organic compounds related to CCS and atmospheric chemistry.

2. Methods

The flow chart for HLC prediction is depicted in Figure 1, which outlines the process of data collection and the conversion of compounds into SMILES, followed by the calculation of molecular descriptors. Feature selection techniques were then applied, and the model was trained on the optimized set of molecular descriptors. Finally, the model was used to predict the HLC of compounds with unknown values.

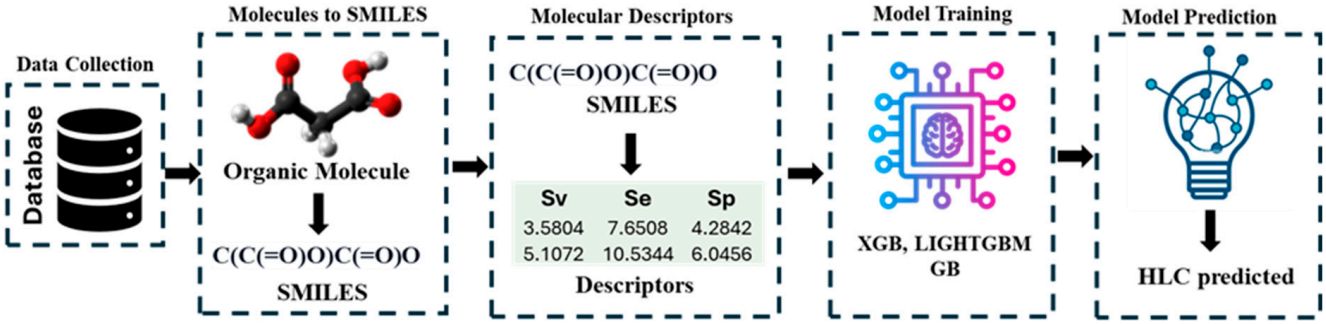


Figure 1. The machine learning approach adopted in this study.

2.1. Dataset

The HLC database of Sander (2023) [19] was used to prepare the dataset for this study. The database assigns an attribute type to each HLC constant, detailing the process of obtaining the constant. To develop a machine learning model, the use of experimental data is recommended. Approximately 1100 experimental HLCs were extracted from the database. The dataset included HLCs for various organic compounds of atmospheric importance, such as alkanes, alkenes, alkynes, amines, amides, aromatics, esters, and ethers. The HLCs obtained from the dataset were Henry's law solubility constants, which were converted into logarithmic form ($\log K_H$).

2.2. Molecular Descriptor Calculation

Molecular descriptors are mathematical representations of molecular properties that play a vital role in the scientific understanding of these properties and behaviors. We calculated the molecular descriptors using alvaDesc software [20]. A list of the SMILES strings of the molecules was generated using CIRPY v1.0.2 software and processed through alvaDesc, calculating all available 2D descriptors. alvaDesc software calculated 4179 descriptors, and these 2D descriptors were grouped into different blocks, such as topological indices, geometrical indices, constitutional indices, and molecular properties. The calculated descriptors and target variables were saved in an Excel file for further data preprocessing.

2.3. Data Preprocessing and Feature Selection

The dataset containing the input features that included molecular descriptors and HLC values of the selected compound was analyzed for any missing and duplicate values. The dataset was preprocessed to handle missing, constant, and duplicate values. Features with low variance were removed using the Sklearn variance threshold feature selector. After applying the variance threshold, a correlation approach was used to remove the highly correlated features.

To address the high dimensionality of our dataset and enhance model interpretability, we adopted a comprehensive feature selection approach that integrated both filter and em-

bedded approaches. Initially, we employed the filter techniques of f_regression and mutual information regression from SelectKBest. These approaches assess the characteristics of a dataset without considering a particular model and determine whether properties have statistically significant relationships (using f_regression) or mutual information (using mutual info regression) with the target variable. We employed a value of k = 50 to determine the most significant characteristics by evaluating their scores using these filter approaches. In addition, we utilized feature importance scores acquired during model training utilizing embedded approaches. We employed LASSO regression and a GB model. LASSO regression is a method that is included in the training process and uses L1 regularization to choose relevant features. It does this by reducing the coefficients of unimportant characteristics to zero. GB models intrinsically assign relevance scores to features depending on their contribution to the model's predictions, serving as an embedded feature selection strategy. This approach led to a conclusive collection of features that included both statistical significance from the filter methods and model-based significance from the embedding approaches. Importantly, the features were evaluated in the context of our domain knowledge. Features with high scores and importance that were also supported by our scientific understanding were retained. The selected features were then used for model development.

2.4. Model Development and Evaluation

The initial dataset was randomly partitioned into separate training, validation, and test datasets to mitigate the risk of data leakage. We used 70% of the original dataset as the training dataset to develop the machine learning models. The validation dataset, comprising 15% of the data, was designated to evaluate the model's fit and assist in hyperparameter tuning. The remaining 15% of the data, which was not used during the model development process, was reserved as the test dataset to evaluate the generalization ability of the developed model.

Five ensemble learning algorithms: AdaBoost, GB, XGB, LightGBM, and CatBoost, along with linear models like Lasso and Elastic Net and tree-based models such as RF, Extra Trees, and DT, were applied to the training datasets using default hyperparameters. The models used in this study are well explained in the literature [21–24]. After identifying the best model, its hyperparameters were further optimized. Bayesian optimization (BO) was used to tune the model's hyperparameters, with RMSE minimization set as the optimization criterion.

During model development, the best feature subset was selected based on the results of feature selection and domain knowledge. The model was retrained on the selected features, and to assess its performance, BO was applied again to optimize the hyperparameter combinations. Ultimately, the optimal machine learning model was obtained by retraining selected features using the most effective combination of hyperparameters. The extent to which the best models can be generalized was further assessed by evaluating their predicted accuracy on a testing dataset, which consisted of unseen data for the model. Subsequently, the most effective models were used to investigate the significance and influence of each feature on the target variable. Finally, the performance of the regression model was assessed using the standard deviation (SD), RMSE, and R^2.

2.5. Feature Importance and Interpretation

The contribution of features to predicting the target variable was evaluated using the SHapley Additive exPlanations (SHAP). SHAP utilizes coalition game theory to compute the Shapley value, aiming to elucidate the prediction model by determining each feature's contribution to the prediction. Shapley values represent the weighted average of the marginal contributions of features over a subset of all feature combinations. They provide a fair estimate of the contribution of distinct features in the sample to prediction outcomes [25].

We used Python (v 3.11.3) for data preprocessing, analysis, ML model development, and interpretation. Model development and computations were conducted on a PC with an Intel® Core™ i5-7500 CPU (Santa Clara, CA, USA) and 32 GB of RAM.

3. Results and Discussion

3.1. Statistical Assessment of the Dataset and Feature Selection

We preprocessed the dataset, which includes the target variable log K_H and molecular descriptors as input features. The dataset initially had 1440 input features. After removing duplicate columns, the number of input features was reduced to 1398. Columns with zero values were dropped to simplify the dataset. Features with low variance were removed using the Sklearn variance threshold feature selector. After dropping features with a threshold variance of 0.05, the number of columns in the dataset was reduced to 1394. Subsequently, the preprocessing steps further reduced the number of input features to 410. Table 1 presents an overview of the summary statistics of the selected input features. Since the dataset contains molecular descriptors with characteristic values for every compound, the input features exhibit diverse ranges for minimum and maximum values. An overview of the range and spread of these variables could aid in examining and analyzing the dataset.

Table 1. Dataset descriptive statistics.

Descriptor	Mean	SD	Maximum	Minimum	Description
MW	110.34	38.92	284.76	26.04	molecular weight
AMW	7.11	3.57	30.76	3.76	average molecular weight
S_v	9.41	3.24	22.92	2.24	sum of the atomic van der Waals volumes
S_e	16.82	6.08	44.35	3.88	sum of the atomic Sanderson electronegativities
S_p	10.16	3.69	26.42	2.21	sum of the atomic polarizabilities
S_i	19.22	7.28	52.72	4.41	sum of the first ionization potentials
M_v	0.57	0.09	1.07	0.43	mean atomic van der Waals volume
M_e	1.00	0.046	1.21	0.95	mean atomic Sanderson electronegativity
M_p	0.61	0.098	1.19	0.50	mean atomic polarizability
M_i	1.14	0.020	1.19	1.07	mean first ionization potential

Additionally, pair-wise correlation analysis was conducted to identify highly correlated features in the dataset. Features with a correlation coefficient greater than 0.95 were removed from the dataset, reducing the number of features to 410. Figure 2 displays the correlation heatmap for the dataset.

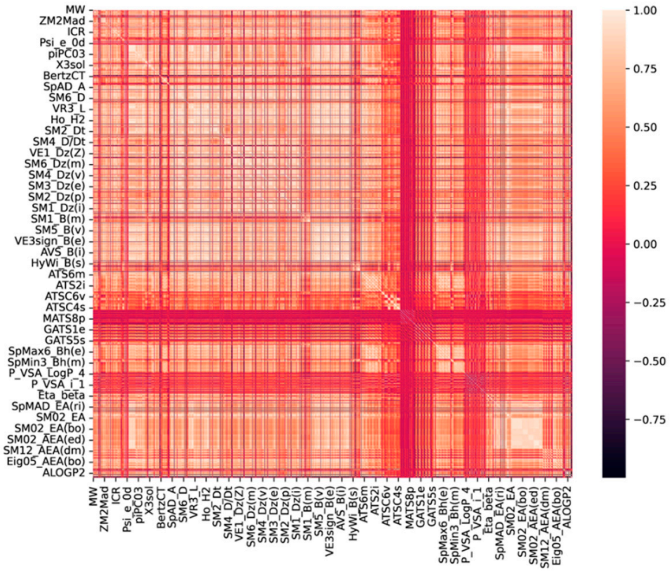


Figure 2. Correlation heatmap of the input features.

Figure 3 displays box plots for selected features, indicating a fairly distributed pattern between the target variable and input features. Some features exhibit a skewed distribution, which can be attributed to the characteristic values of compounds in the dataset since different molecules have varying values for specific features. Histogram plots illustrating the data distribution of the selected features and target variables are presented in Figure S1. Additionally, Figure S2 depicts plots for selected pairs of features that have a correlation higher than 0.9.

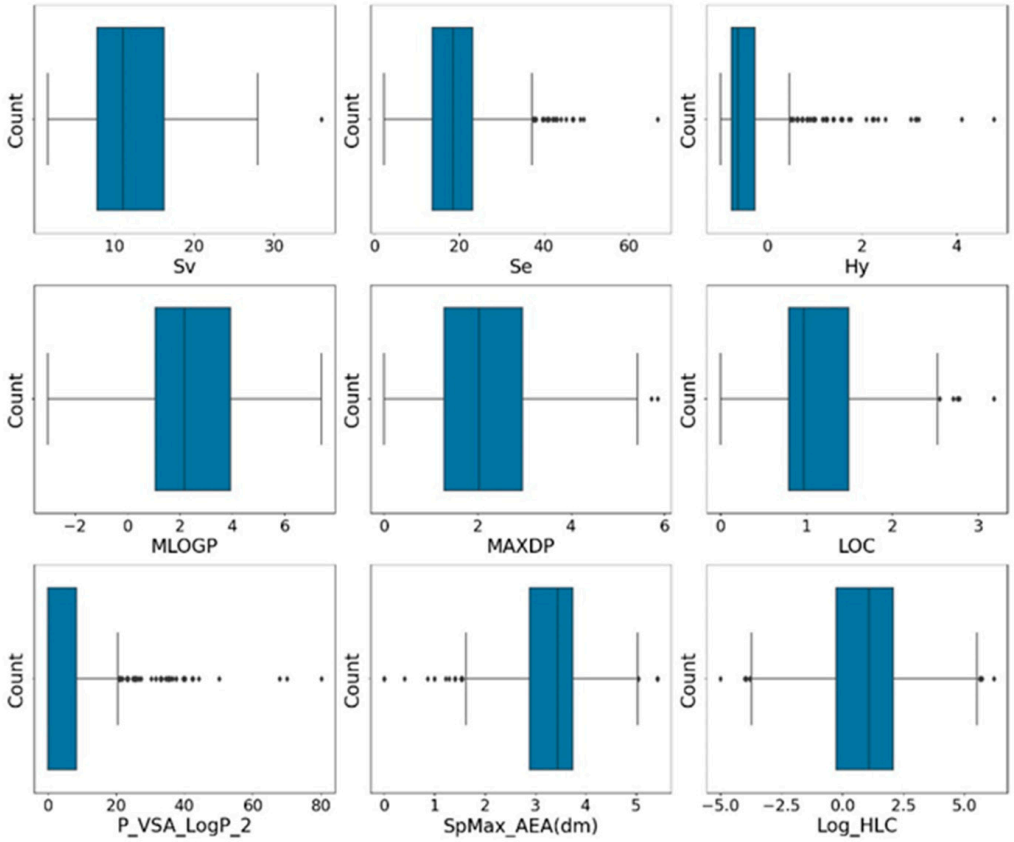


Figure 3. Box plots for selected features and target variables of log HLC.

Figure 4 presents charts of features selected based on F_regression, mutual information, and importance from the model. Figure 4a,b display the features selected based on the importance ranking provided by GB and Lasso. In Figure 4a, the descriptors P_VSA_i_2, P_VSA_s_6, and P_VSA_p_1, with importance scores of 0.025, 0.016, and 0.0145, respectively, demonstrate their contribution to model accuracy. Figure 4b illustrates that P_VSA_p_2, followed by GATS1e, have importance scores of 0.38 and 0.075, respectively. In Figure 4c, the selected features demonstrate a strong linear relationship with the target variable. Conversely, in Figure 4d, the features selected using the mutual information method capture the underlying nonlinearity between the features and the target variable. The models were then evaluated using the feature subset obtained from the feature selection method, focusing on the features that significantly contribute to model performance. These features were also assessed based on the domain knowledge and their relationship to the target variable. Table S1 lists the best feature subset used in the final model training.

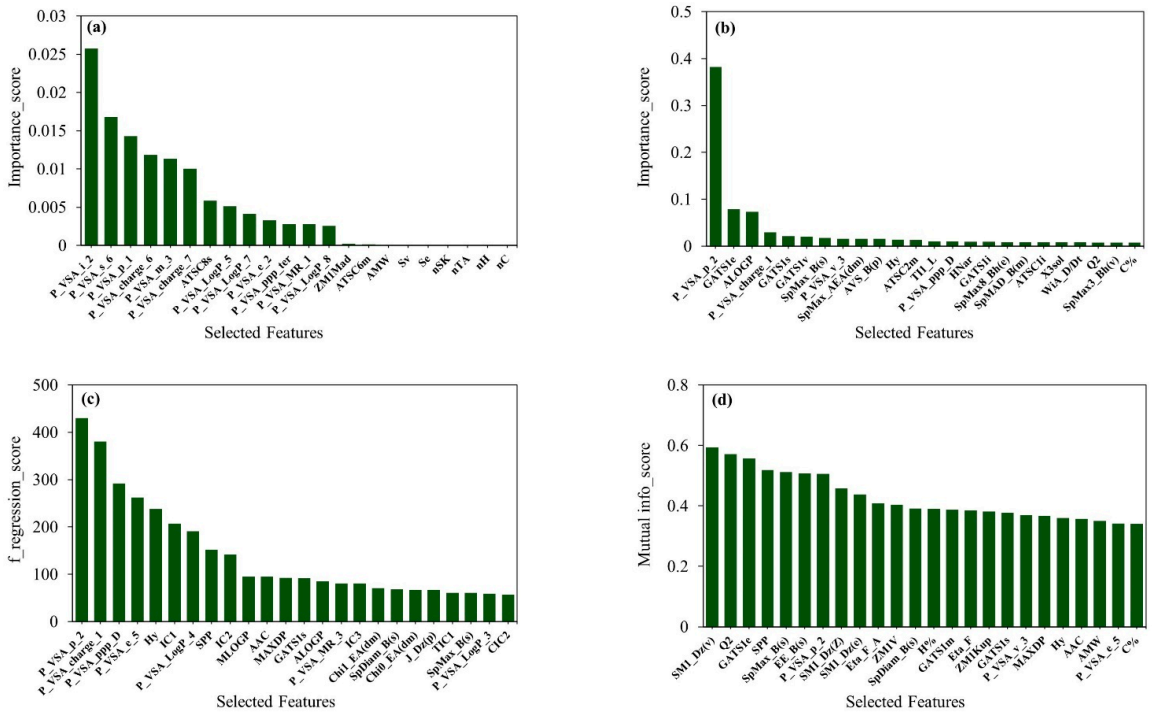


Figure 4. Feature selection (**a**) from the model (GB), (**b**) from the model (Lasso), (**c**) F_regression, and (**d**) mutual information regression.

3.2. Model Development and Selection

Various machine learning algorithms were evaluated to address the challenges presented by complex datasets. Cross-validation and Bayesian optimization techniques were employed to enhance the preferred algorithms, resulting in robust predictive models capable of generalizing well on unseen data. Model performance was evaluated in terms of normalized RMSE using a Taylor diagram, as depicted in Figure 5. According to Nguyen et al. (2021), there was a positive correlation between the proximity of a point to a reference point (REF) and its performance [26]. The results suggested that the boosting models XGBoost, GB, CatBoost, and LightGBM exhibited high accuracy. However, the values for the Lars and Elastic Net models are not shown in Figure 5, because their RMSE values are very high.

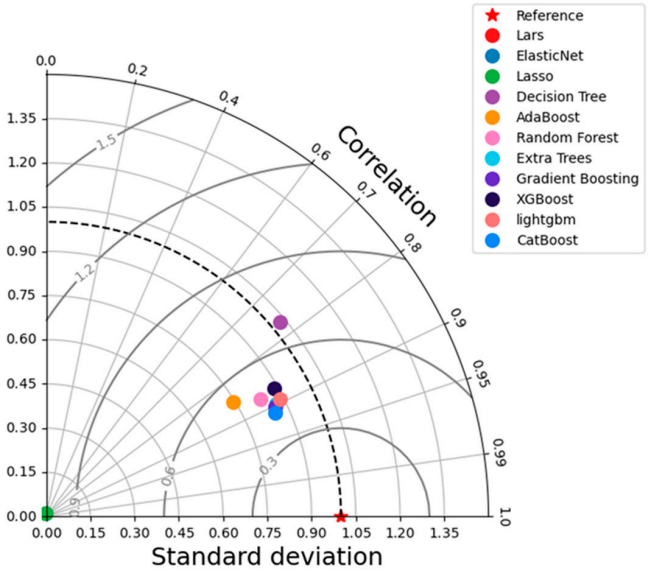


Figure 5. Taylor diagram of the algorithms used in this study.

Table 2 summarizes the statistical metrics for the machine learning models used to predict HLC. The metrics used in this study are derived from the mean value obtained from five-fold cross-validation for both the training and validation datasets. It can be inferred from Table 2 that the Extra Trees perform well on the training data, achieving a RMSE of 0.0518 and R^2 of 0.9993, followed by XGB and CatBoost with RMSE values of 0.1050 and 0.0520, respectively. Conversely, when applied to the test data, the GB model exhibited superior performance compared to CatBoost and XGB, with a validation RMSE of 0.8275. This can be attributed to the higher complexity of GB, which enables it to handle complex datasets. Gradient Boosting is renowned for its ability to capture intricate nonlinear correlations within the data and generate predictions with exceptional accuracy. Variations in the scale of different features do not affect the algorithm. Its main objective is to optimize the loss function through gradient descent rather than depending on feature scaling.

Table 2. Statistical metrics for the proposed machine learning models.

Model	R^2_Train	RMSE_Train	R^2_Validation	RMSE_Validation
Extra Trees	0.9993	0.0518	0.8383	0.8137
Gradient Boosting	0.9584	0.4901	0.8328	0.8275
CatBoost	0.9970	0.1050	0.8283	0.8386
Random Forest	0.9720	0.3205	0.8179	0.8635
XGBoost	0.9993	0.0520	0.8131	0.8749
LightGBM	0.9959	0.1225	0.8090	0.8843
AdaBoost	0.8046	0.8459	0.7191	1.0725

Furthermore, the models were evaluated based on the selected features using various feature selection methods. This provides insight into the performance of different models in relation to the RMSE. Table 3 presents the RMSE for the models trained on the subset of features, and it was inferred that GB, CatBoost, and LightGBM yielded lower RMSE compared to AdaBoost, RF, and DT.

Table 3. RMSE of models trained on selected features.

Feature Selection Technique	AdaBoost	CatBoost	Decision Tree	Extra Trees	Gradient Boosting	Light GBM	Random Forest	XG Boost
F_regression	0.764	0.622	0.888	0.590	0.659	0.697	0.702	0.735
Mutual_info_regression	0.802	0.637	0.946	0.606	0.690	0.674	0.800	0.810
SelectfromModel_Lasso	0.855	0.668	1.046	0.593	0.700	0.751	0.786	0.745
SelectfromModel_GB	0.708	0.561	0.882	0.554	0.543	0.619	0.654	0.617

3.3. Model Prediction

Based on the evaluation metrics in the model selection, GB was selected for the prediction task. Bayesian auto-optimization was employed, resulting in the final model, specifically the Bayesian-optimized GB model. The optimized parameters for GB are displayed in Table 4. We then retrained the optimized model on the training dataset to assess its performance on the test data.

Table 4. Optimized hyperparameters for GB.

Parameter	Search Space	Optimum Values
learning_rate	(0.005, 1.0)	0.134
n_estimators	(5, 200)	193
min_samples_split	(2, 100)	44
min_samples_leaf	(1, 100)	59
Max_depth	(2, 50)	27

To thoroughly evaluate the applicability of our final model, we utilized two separate evaluation sets: a validation set and a test set. Although the R^2 score obtained on the training set was 0.96, showing a high level of correlation between the predicted and actual values, we employed a separate validation dataset to verify the model's capacity to generalize to new data. The R^2 value for the validation set was 0.78, indicating a high degree of generalizability. This observation was further corroborated by the R^2 score of 0.74 obtained on the independent test set. Figure 6b illustrates the result of the model performance evaluation, indicating that the final optimal model achieves high prediction accuracy for HLC with a R^2 value of 0.74. The correlation between experimental HLC values and model predictions is clearly demonstrated in Figure 6b. However, some deviation exists, which may be attributed to variability within the data and the intrinsic complexity of chemical compounds within the extensive dataset. The R^2 value of 0.74 implies that the model successfully encapsulates a significant part of the fundamental data. Within the context of generalization, a R^2 value of this scale on the test data is significant, because it indicates that the model not only fits the training data effectively but also generalizes well to new data. Furthermore, the model's ability to generalize serves as a reliable indicator of its usefulness in practical applications, such as in generating HLC predictions. If the model demonstrates a satisfactory level of accuracy on the test data, it indicates that it can be reliably used to make predictions for new, unseen data.

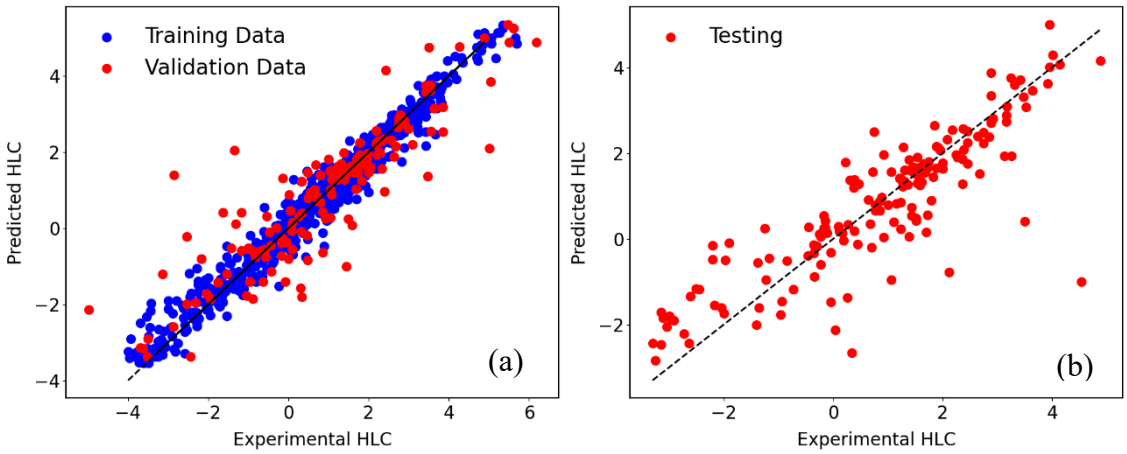


Figure 6. BO–GB model performance. (**a**) Training and validation set and (**b**) testing set.

3.4. Model Interpretation

An efficiently trained model is capable of predicting HLC, while SHAP offers an interpretation of the ML models. SHAP explains the contribution of features to the model's prediction ability. Figures 7 and 8 depict the ranking of descriptors according to their contributions to the output variables. Figure 7 provides qualitative information on the effect of the variable on the target variable, whereas Figure 8 provides a quantitative interpretation of the correlation between the feature and target value.

As depicted in Figure 8, *P_VSA_e_5*, which belongs to the P_VSA-like descriptor block representing Sanderson electronegativity, ranks higher, with a Shapley value of 0.4392. The correlation between HLC and electronegativity is illustrated by the blue markers, indicating that a high *P_VSA_e_5* value corresponds to a low HLC volatility constant, whereas a negative value signifies high HLC volatility. The second variable correlating with the target variable HLC is *P_VSA_charge_1*, which has a Shapley value of 0.3597, as shown in Figure 8. This descriptor describes strong intermolecular interactions based on the charged partial surface area [27]. As illustrated in Figure 7, low-value *P_VSA_charge_1* exhibits a negative correlation with the target variable HLC, because the presence of partial charges in the molecules leads to strong water interaction through hydrogen bonding. The Moriguchi octanol–water partition coefficient (*MLOGP*), which indicates the partitioning

behavior of chemical compounds, is shown in Figure 8 with a SHAP value of 0.3917. The charge descriptor *Q2*, representing the total squared charge, explains the distribution of electronic behavior within a molecule and has a Shapley value of 0.3017. This high value underscores its impact on the model-predicted HLC values. The relationship between *Q2* and *log P* has been explored in the literature [28], concluding that *Q2* is dependent on the molecular structure and its charge distribution.

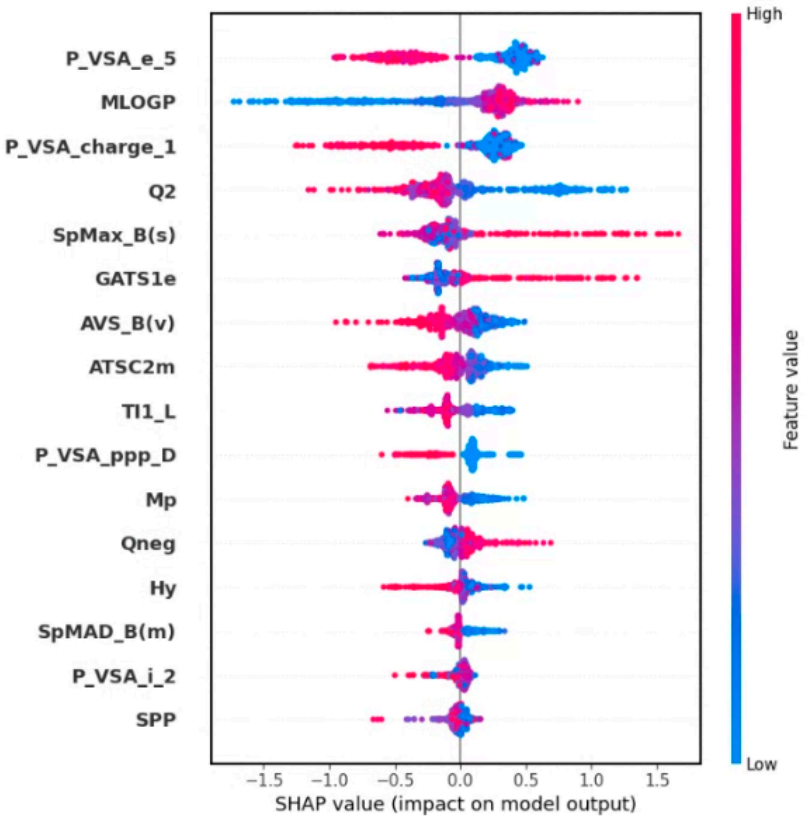


Figure 7. SHAP summary plot for the BO–GB model.

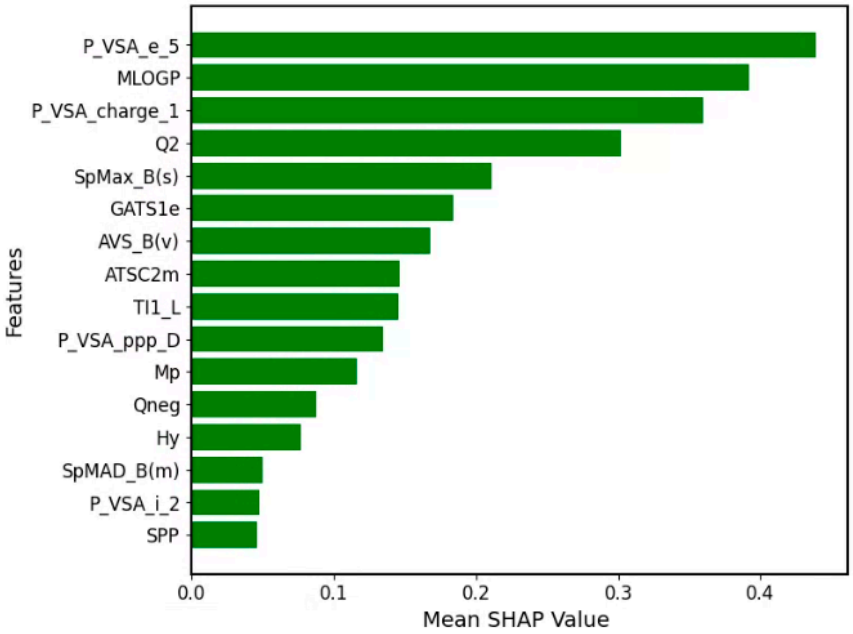


Figure 8. SHAP feature importance measured as the mean absolute Shapley values.

The average vertex sum from the Burden matrix, weighted by a van der Waals volume *(AVS_B(v))* with a Shapley value of 0.1676, underscores its impact on the model prediction. Figure 7 illustrates that a high value of this feature negatively affects HLC. This can be explained in terms of the atomic van der Waal's volume distribution of atoms in a molecule; molecules with a bulky or branched structure interact more with solute molecules, thereby resulting in a low Henry's law constant of volatility. In addition to the charge distribution, the mass distribution within a molecule influences its physical properties. The Special Mean Absolute Deviation of Burden matrix weighted by mass (*SpMAD_B(m)*), with a Shapley value of 0.05, demonstrates its impact on the target variable. As depicted in Figure 7, a high value negatively impacts the target variable, indicating low volatility for compounds with a high mass distribution within a molecule.

GATS1e is defined as the Geary autocorrelation coefficient lag1 weighted by Sanderson electronegativity. This descriptor serves as an indicator of the polar interactions within a molecule. As depicted in Figure 7, it exerts a positive impact on the model output when its value is high. Conversely, the molecular descriptor *Hy*, which represents the hydrophilic properties of molecules, negatively affects the model output. This indicates that molecules with high hydrophilic values lead to a decrease in the Henry's law constant of volatility. The model also integrates a parameter known as *Qneg*, representing the total negative charge, which positively influences the model. *P_VSA_ppp_D*, a P_VSA-like descriptor for potential pharmacophore points, specifically a D-hydrogen bond donor, has a Shapley value of 0.1346, demonstrating its contribution to the model's prediction ability. The SHAP dependence plots of the input features are depicted in Figure S3.

3.5. Model Comparison and Application

In this study, the predictive accuracy of the developed ML model was compared to that of traditional models that use vapor pressure and aqueous solubility as the key variables for HLC estimation. As depicted in Figure 9, there exists a significant correlation of 0.80 between the HLC values predicted by the ML model and those estimated by the traditional vapor pressure–aqueous solubility model. While there is a strong general agreement between the two models, the study identified approximately 10% of the examined compounds as outliers. This discrepancy primarily stems from the traditional model's dependence on aqueous solubility values. It is crucial to underscore that the solubility of a substance in water is not an intrinsic property but rather hinges on the specific measurement methods employed. Differences in these approaches can introduce inconsistencies in the solubility measurements, thereby affecting the reliability of the predictions provided by the vapor pressure–aqueous solubility model.

A dearth of comprehensive models exist that can accurately estimate the K_H values for many types of substances. Models reported in the scientific literature are typically designed for a specific group or a category of substances. For instance, the models developed by Modarresi et al. (2005), Duchowicz et al. (2008), and Goodarz et al. (2010) were employed for specific classes of organic compounds [29–31]. Few models that accommodate a wider range of compounds require the use of the three-dimensional molecular structure, necessitating a complex computation to determine the most stable conformer [32–34]. Our model boasts the advantage of being applicable to a diverse array of compounds with varying characteristics. Its user-friendliness stems from its simplicity, eliminating the need to determine or compute the 3D structures of the compounds. Furthermore, the model presented in our study surpasses previous models, because it was developed using a wide range of experimental HLC values from approximately 1100 pure compounds across various chemical families. Nevertheless, we evaluated our model by comparing it with a reliable approach based on aqueous solubility and vapor pressure. To evaluate the predictive accuracy of our model, we employed compounds that were not part of the model's development. The experimental K_H values of these compounds are not known. The values predicted by the model were compared to those estimated by the aqueous solubility and vapor pressure method. In conclusion, our developed model provides several clear advan-

tages. First and foremost, the development of this model is based on reliable experimental data, thereby improving reliability in its predictions. Furthermore, the training dataset includes an extensive variety of chemical compounds, ranging from basic molecules such as ethane to intricate structures like borneol ((2*R*)-1,7,7-trimethylbicyclo [2.2.1] heptan-2-ol) and hexazinone (3-cyclohexyl-6-(dimethylamino)-1-methyl-1,3,5-triazine-2,4-dione). The vast coverage of this data ensures that it may be applied to a wide range of scientific purposes with high accuracy and reliability. Finally, the model leverages the simplicity of 2D molecular descriptors, which not only facilitates interpretability but also enhances its computational efficiency.

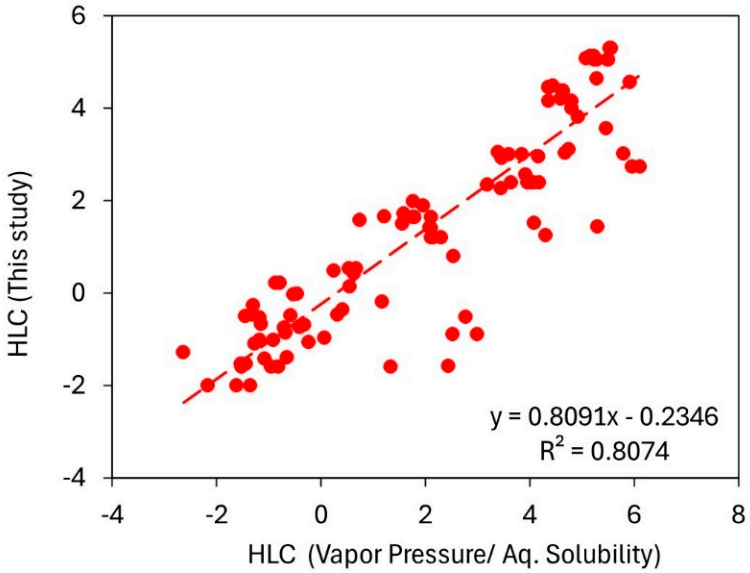


Figure 9. Correlation of HLC values predicted by the vapor pressure–aqueous solubility model and the ML model.

We applied the developed model to predict the HLC of compounds related to CCS and SOA. Amines and their degradation products from CCS, along with dicarboxylic acid related to SOA, were sourced from the literature [35,36]. The results of the predicted HLC values are presented in Table 5. We evaluated the predicted values for the compounds by comparing their structural similarity with compounds of known HLC values. For example, the experimentally measured HLC value for 2-amino-2-methyl-1-propanol (2-AMP) is 4.76×10^{-3} m^3 Pa mol^{-1}, while the values predicted by the current model for the structurally similar compounds like 2-butylamino ethanol (2-BAE) and 2-amino-1-butanol (2-AB) are 2.66×10^{-1} m^3 Pa mol^{-1} and 3.10×10^{-2} m^3 Pa mol^{-1}, respectively. Similarly, the experimentally determined HLC value for piperazine is 2.22×10^{-4} m^3 Pa mol^{-1}, while the ML model predicted values for its derivatives of 1,4-dimethyl piperazine (1,4 DMPZ) and *N*-ethylpiperazine are 2.21×10^{-1} m^3 Pa mol^{-1} and 1.78×10^{-1} m^3 Pa mol^{-1}, respectively. In the case of N-methyl cyclohexylamine, the ML model's predicted value of 1.08 m^3 Pa mol^{-1} aligns well with the HLC value of its structurally similar compound *N*,*N*-dimethylcyclohexylamine at 2.38 m^3 Pa mol^{-1}. The HLC values predicted by the ML model align well with the HLC values of their structurally relevant compounds, consistent with the underlying scientific concepts. Generally, the degree of alkylation decreases the compound's water solubility, thereby increasing its HLC value of volatility. This trend is exemplified by piperazine and its alkyl derivatives, where the latter's HLC values surpass the former's. This trend is also evident in the alkanol amine compounds, leading to the conclusion that the model can reliably predict the HLC values of compounds associated with CCS.

Table 5. Predicted value of HLC for non-measured compounds.

Compounds	Formula	H_s^{cp} m^3 Pa mol^{-1}
2-(Butylamino)ethanol	$C_6H_{15}NO$	2.66×10^{-1}
2-Amino-1-butanol	$C_6H_{11}NO$	3.10×10^{-2}
N-(Hydroxyethyl)ethylenediamine	$C_6H_{12}N_2O$	1.33×10^{-2}
N-(2-Hydroxyethyl)acetamide	$C_6H_9NO_2$	1.33×10^{-2}
1,4-Dimethyl piperazine	$C_6H_{14}N_2$	2.28×10^{-1}
N-(2-hydroxyethyl)formamide	$C_6H_7NO_2$	1.33×10^{-2}
1-(2-Hydroxyethyl)imidazole	$C_5H_8N_2O$	1.13×10^{-2}
2-Hydroxy-N-(2-hydroxyethyl)acetamide	$C_4H_9NO_3$	3.28×10^{-3}
N-(2-hydroxyethyl)oxamic acid	$C_4H_7NO_4$	6.70×10^{-3}
N,*N*′-bis(2-hydroxyethyl)oxamide	$C_6H_{12}N_2O_4$	3.63×10^{-3}
N-formylpiperazine	$C_5H_{10}N_2O$	3.28×10^{-3}
N-ethylpiperazine	$C_6H_{14}N_2$	1.78×10^{-1}
4-diethylamino-2-butanol	$C_8H_{19}NO$	2.63×10^{-1}
N,*N*,*N*′,*N*′-tetramethylpropane-1,3-diamine	$C_7H_{18}N_2$	3.59×10^{-1}
1-(dimethylamino)propan-2-ol	$C_5H_{13}NO$	4.78×10^{-2}
N-methylcyclohexylamine	$C_7H_{15}N$	1.08×10^{0}
Dodecanedioic acid	$C_{12}H_{22}O_4$	1.23×10^{-1}
Methylmalonic acid	$C_4H_6O_4$	1.23×10^{-1}
2,2-Dimethylmalonic acid	$C_5H_8O_4$	6.70×10^{-3}
Ethyl malonic acid	$C_5H_8O_4$	1.05×10^{-2}
Butyl malonic acid	$C_7H_{12}O_4$	1.05×10^{-2}
Methylsuccinic acid	$C_5H_8O_4$	9.38×10^{-2}
2,2-Dimethylsuccinic acid	$C_6H_{10}O_4$	1.05×10^{-2}
2-Methylglutaric acid	$C_6H_{10}O_4$	3.02×10^{-2}
2,2-Dimethylglutaric acid	$C_7H_{12}O_4$	3.02×10^{-2}
3,3-Dimethylglutaric acid	$C_7H_{12}O_4$	9.38×10^{-2}
aspartic acid	$C_4H_7NO_4$	3.02×10^{-2}
2-Oxosuccinic acid	$C_4H_4O_5$	9.38×10^{-2}
3-Oxoglutaric acid	$C_5H_6O_5$	6.70×10^{-3}
2-Oxoadipic acid	$C_6H_8O_5$	6.70×10^{-3}
4-Oxopimelic acid	$C_7H_{10}O_5$	4.41×10^{-2}
2-hydroxymalonic acid	$C_3H_4O_5$	1.07×10^{-2}
1,2 Cyclopentanedicarboxylic acid	$C_7H_{10}O_4$	3.88×10^{-2}

We also applied the ML model to predict the HLC values of dicarboxylic acids related to SOA. In the case of malonic acid, the degree of substitution ranges from methyl to butyl groups. The model's predicted HLC values for the alkyl-substituted malonic acid range from 6.70×10^{-3} m^3 Pa mol^{-1} to 9.38×10^{-2} m^3 Pa mol^{-1}. This increase in HLC, indicative of volatilization, indicates the lower solubility of the substituted carboxylic acids with an increase in the alkyl chain. The same trend was observed for alkyl-substituted succinic acid and glutaric acid, as shown in Table 5. Consequently, it can be concluded that the HLC values predicted by the model for the dicarboxylic acids associated with SOA are reliable and can be used in atmospheric modeling studies.

4. Conclusions

This study developed a machine learning model using experimental HLC data. Molecular descriptors were employed as input features, and a boosting algorithm was used for model training. Various feature selection methods generated a subset of features that translated the underlying chemistry of HLC. Training and validation of the models were conducted using the training and validation datasets. The model based on the GB algorithm exhibited satisfactory performance. The hyperparameters of the GB model were optimized using an automated optimization framework known as Optuna. The optimized model underwent another round of training on the training data and was subsequently evaluated

using a separate test dataset. The chosen metrics for evaluating model performance were RMSE and R^2. A R^2 value of 0.75 indicated satisfactory model performance. The accuracy of the model's predictions was evaluated using an external dataset of 36 organic compounds related to CCS and SOA. This evaluation confirms that the model can be used to estimate HLC values for compounds with unknown partition coefficients. The complexity of the dataset and the large variation in the molecular structure could be the underlying factors influencing the model's R^2 value. However, the model's ability to accurately represent the complex characteristics of the molecules, along with its user-friendly interface compared to the models developed by [11,12], demonstrates its reliability and robustness. Future enhancements to the model will include the incorporation of 3D molecular descriptors and an update for predicting vapor pressure data using the predicted HLC.

Future research will adopt two main approaches to enhance the ML model outlined in this paper. The first approach involves incorporating more molecules to expand the model's range of applications, and the second entails exploring new molecular descriptors to provide more comprehensive insights into molecular structures. These improvements will not only augment the model's predictive accuracy, particularly for organic chemicals, but also incorporate updates for predicting vapor pressure data using the projected Henry's law constants (HLC). To facilitate the accessibility and application of these precise predictions, a web-based graphical user interface (GUI) will be developed.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/atmos15060706/s1, Figure S1: Histograms of selected input features and target variable; Figure S2: Pair Plots of highly correlated features; Figure S3: SHAP dependence plots of model input features; Table S1: Features (molecular descriptors) for model development.

Author Contributions: A.U.: Investigation, Software, Methodology, Validation, Data Curation, and Writing—Original Draft. M.S.: Conceptualization and Software. H.-J.L.: Writing—Review and Editing and Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to, the data used in this study are reserved for further analysis by the authors.

Code Availability: The code used in this study is available at the following repository link: https://github.com/Atta1989/Henry-Law-Constant-Prediction (accessed on 5 June 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mackay, D.; Boethling, R.S. Henry's Law Constant. In *Handbook of Property Estimation Methods for Chemicals*; CRC Press: Boca Raton, FL, USA, 2000; pp. 91–110. ISBN 0429133294.
2. Li, M.; Wang, X.; Zhao, Y.; Du, P.; Li, H.; Li, J.; Shen, H.; Liu, Z.; Jiang, Y.; Chen, J.; et al. Atmospheric Nitrated Phenolic Compounds in Particle, Gaseous, and Aqueous Phases during Cloud Events at a Mountain Site in North China: Distribution Characteristics and Aqueous-Phase Formation. *J. Geophys. Res. Atmos.* **2022**, *127*, e2022JD037130. [CrossRef]
3. Xuan, X.; Chen, Z.; Gong, Y.; Shen, H.; Chen, S. Partitioning of Hydrogen Peroxide in Gas-Liquid and Gas-Aerosol Phases. *Atmos. Chem. Phys.* **2020**, *20*, 5513–5526. [CrossRef]
4. Leng, C.; Kish, J.D.; Roberts, J.E.; Dwebi, I.; Chon, N.; Liu, Y. Temperature-Dependent Henry's Law Constants of Atmospheric Amines. *J. Phys. Chem. A* **2015**, *119*, 8884–8891. [CrossRef] [PubMed]
5. Staudinger, J.; Roberts, P. V A Critical Review of Henry's Law Constants for Environmental Applications. *Crit. Rev. Environ. Sci. Technol.* **1996**, *26*, 205–297. [CrossRef]
6. Linnemann, M.; Nikolaychuk, P.A.; Muñoz-Muñoz, Y.M.; Baumhögger, E.; Vrabec, J. Henry's Law Constant of Noble Gases in Water, Methanol, Ethanol, and Isopropanol by Experiment and Molecular Simulation. *J. Chem. Eng. Data* **2020**, *65*, 1180–1188. [CrossRef]

7. Keshavarz, M.H.; Rezaei, M.; Hosseini, S.H. A Simple Approach for Prediction of Henry's Law Constant of Pesticides, Solvents, Aromatic Hydrocarbons, and Persistent Pollutants without Using Complex Computer Codes and Descriptors. *Process Saf. Environ. Prot.* **2022**, *162*, 867–877. [CrossRef]
8. Nirmalakhandan, N.N.; Speece, R.E. QSAR Model for Predicting Henry's Constant. *Environ. Sci. Technol.* **1988**, *22*, 1349–1357. [CrossRef]
9. Meylan, W.M.; Howard, P.H. Bond Contribution Method for Estimating Henry's Law Constants. *Environ. Toxicol. Chem.* **1991**, *10*, 1283–1293. [CrossRef]
10. Lin, S.T.; Sandler, S.I. Henry's Law Constant of Organic Compounds in Water from a Group Contribution Model with Multipole Corrections. *Chem. Eng. Sci.* **2002**, *57*, 2727–2733. [CrossRef]
11. Gharagheizi, F.; Abbasi, R.; Tirandazi, B. Prediction of Henry's Law Constant of Organic Compounds in Water from a New Group-Contribution-Based Model. *Ind. Eng. Chem. Res.* **2010**, *49*, 10149–10152. [CrossRef]
12. Duchowicz, P.R.; Aranda, J.F.; Bacelo, D.E.; Fioressi, S.E. QSPR Study of the Henry's Law Constant for Heterogeneous Compounds. *Chem. Eng. Res. Des.* **2020**, *154*, 115–121. [CrossRef]
13. Vo Thanh, H.; Zhang, H.; Dai, Z.; Zhang, T.; Tangparitkul, S.; Min, B. Data-Driven Machine Learning Models for the Prediction of Hydrogen Solubility in Aqueous Systems of Varying Salinity: Implications for Underground Hydrogen Storage. *Int. J. Hydrogen Energy* **2024**, *55*, 1422–1433. [CrossRef]
14. Hou, Y.; Wang, S.; Bai, B.; Stephen Chan, H.C.; Yuan, S. Accurate Physical Property Predictions via Deep Learning. *Molecules* **2022**, *27*, 1668. [CrossRef] [PubMed]
15. Zhang, W.; Wang, Y.; Ren, S.; Hou, Y.; Wu, W. Novel Strategy of Machine Learning for Predicting Henry's Law Constants of CO_2 in Ionic Liquids. *ACS Sustain. Chem. Eng.* **2023**, *11*, 6090–6099. [CrossRef]
16. Wu, T.; Li, W.L.; Chen, M.Y.; Zhou, Y.M.; Zhang, Q.Y. Prediction of Henry's Law Constants of CO_2 in Imidazole Ionic Liquids Using Machine Learning Methods Based on Empirical Descriptors. *Chem. Pap.* **2021**, *75*, 1619–1628. [CrossRef]
17. Orhan, I.B.; Le, T.C.; Babarao, R.; Thornton, A.W. Accelerating the Prediction of CO_2 Capture at Low Partial Pressures in Metal-Organic Frameworks Using New Machine Learning Descriptors. *Commun. Chem.* **2023**, *6*, 214. [CrossRef] [PubMed]
18. Wang, Q.; Yao, A.; Shokri, M.; Dineva, A.A. Predictive Modeling of Henry's Law Constant in Chemical Structures Using LSSVM and ANFIS Algorithms. *Preprints* **2020**, 2020020248. [CrossRef]
19. Sander, R. Compilation of Henry's Law Constants (Version 5.0.0) for Water as Solvent. *Atmos. Chem. Phys.* **2023**, *23*, 10901–12440. [CrossRef]
20. Mauri, A. AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*; Methods in Pharmacology and Toxicology (MIPT); Springer: Berlin/Heidelberg, Germany, 2020; pp. 801–820. [CrossRef]
21. Lomte, D.R.S.S.; Torambekar, M.R.S.G.; Janwale, M.R.A.P. Methods, Theory of Boosting Algorithm: A Review. *JournalNX* **2018**, 39–44. Available online: https://repo.journalnx.com/index.php/nx/article/view/2024 (accessed on 5 June 2024).
22. Gulati, P.; Sharma, A.; Gupta, M. Theoretical Study of Decision Tree Algorithms to Identify Pivotal Factors for Performance Improvement: A Review. *Int. J. Comput. Appl.* **2016**, *141*, 19–25. [CrossRef]
23. Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2015**, *25*, 197–227. [CrossRef]
24. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A Comparative Analysis of Gradient Boosting Algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
25. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *2017*, 4766–4775.
26. Nguyen, D.H.; Hien Le, X.; Heo, J.-Y.; Bae, D.-H. Development of an Extreme Gradient Boosting Model Integrated with Evolutionary Algorithms for Hourly Water Level Prediction. *IEEE Access* **2021**, *9*, 125853–125867. [CrossRef]
27. Stanton, D.T.; Jurs, P.C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure-Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323–2329. [CrossRef]
28. Karelson, M.; Lobanov, V.S.; Katritzky, A.R. Quantum-Chemical Descriptors in QSAR/QSPR. *Chem. Rev.* **1996**, *96*, 1027–1044. [CrossRef] [PubMed]
29. Modarresi, H.; Modarress, H.; Dearden, J.C. Henry's Law Constant of Hydrocarbons in Air–Water System: The Cavity Ovality Effect on the Non-Electrostatic Contribution Term of Solvation Free Energy. *SAR QSAR Environ. Res.* **2005**, *16*, 461–482. [CrossRef] [PubMed]
30. Goodarzi, M.; Ortiz, E.V.; Coelho, L.d.S.; Duchowicz, P.R. Linear and Non-Linear Relationships Mapping the Henry's Law Parameters of Organic Pesticides. *Atmos. Environ.* **2010**, *44*, 3179–3186. [CrossRef]
31. Duchowicz, P.R.; Garro, J.C.; Castro, E.A. QSPR Study of the Henry's Law Constant for Hydrocarbons. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 133–140. [CrossRef]
32. Gharagheizi, F.; Ilani-Kashkouli, P.; Mirkhani, S.A.; Farahani, N.; Mohammadi, A.H. QSPR Molecular Approach for Estimating Henry's Law Constants of Pure Compounds in Water at Ambient Conditions. *Ind. Eng. Chem. Res.* **2012**, *51*, 4764–4767. [CrossRef]
33. Modarresi, H.; Modarress, H.; Dearden, J.C. QSPR Model of Henry's Law Constant for a Diverse Set of Organic Chemicals Based on Genetic Algorithm-Radial Basis Function Network Approach. *Chemosphere* **2007**, *66*, 2067–2076. [CrossRef] [PubMed]
34. O'Loughlin, D.R.; English, N.J. Prediction of Henry's Law Constants via Group-Specific Quantitative Structure Property Relationships. *Chemosphere* **2015**, *127*, 1–9. [CrossRef] [PubMed]

35. Bilde, M.; Barsanti, K.; Booth, M.; Cappa, C.D.; Donahue, N.M.; Emanuelsson, E.U.; McFiggans, G.; Krieger, U.K.; Marcolli, C.; Topping, D.; et al. Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures. *Chem. Rev.* **2015**, *115*, 4115–4156. [CrossRef] [PubMed]
36. Sharif, M.; Fan, H.; Wu, X.; Yu, Y.; Zhang, T.; Zhang, Z. Assessment of Novel Solvent System for CO_2 Capture Applications. *Fuel* **2023**, *337*, 127218. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.