ChemPlusChem

Research Article
doi.org/10.1002/cplu.202400459

Chemistry Europe
European Chemical Societies Publishing

www.chempluschem.org

A QSPR Model for Henry's Law Constants of Organic Compounds in Water and Ethanol for Distilled Spirits

John White,[a] Johnathan Graf,[a] Samuel Haines,[a] Noppadon Sathitsuksanoh,[a] R. Eric Berson,[a] and Vance W. Jaeger*[a]

Henry's law describes the vapor-liquid equilibrium for dilute gases dissolved in a liquid solvent phase. Descriptions of vapor-liquid equilibrium allow the design of improved separations in the food and beverage industry. The consumer experience of taste and odor are greatly affected by the liquid and vapor phase behavior of organic compounds. This study presents a machine learning (ML) based model that allows quick, accurate predictions of Henry's law constants (k_H) for many common organic compounds. Users input only a Simplified Molecular-Input Line-Entry System (SMILES) string or a common English name, and the model returns Henry's law estimates for compounds in water and ethanol. Training was performed on 5,690 compounds. Training data were gathered from an existing database and were supplemented with quantum mechanical (QM) calculations. An extra trees regression model was generated that predicts k_H with a mean absolute error of 1.3 in log space and an R^2 of 0.98. The model is applied to common flavor and odor compounds in bourbon whiskey as a test case for food and beverage applications.

Introduction

Henry's law states that at a given temperature the amount of gas that dissolves in a liquid is proportional to the partial pressure of the gas in the vapor phase. Henry's law is stated succinctly in Equation (1). C is the concentration of the gas in the liquid in units of mol/L or M for molarity. Henry's law constant is k_H which is expressed as a ratio of the solute phase concentration over the partial pressure. Others often denote a Henry's law constant as H rather than k_H. P is the partial pressure of the gas above the liquid interface in units of atm. Therefore, k_H adopts units of mol/L-atm or M/atm.

$$C = k_H P \quad (1)$$

The value of k_H depends upon the chemical composition of the solvent, the compound being dissolved, and temperature. For the current study, temperature dependence is not considered, and all values are taken at a reference temperature of 298.15 K. Henry's law is most applicable in dilute solutions. Thus, it is well suited for applications in which volatile organic compounds are dissolved in a liquid at low concentrations such as with flavor and odor compounds in distilled spirits, wines, and other beverages. Apart from food and beverage applications, Henry's law is often applied to studies of dilute pollutants in the environment.

Sander has compiled an extensive online database of Henry's law constants of compounds in water.[1] The most recent iteration of this database (version 5.0.0) reportedly contains over 46,000 values of k_H representing over 10,000 species. Dickman recently reported the development of an accurate predictive model for Henry's law constants based upon an earlier version of Sander's database.[2] Dickman compiled data about chemical functional groups and physical characteristics of thousands of molecules to demonstrate the accuracy several ML models for predicting k_H in water. Toropov and coworkers recently established a model relationship between molecular structure (encoded in a SMILES string), molecular properties, and reported values of k_H. The model proved accurate across a large range of chemical species.[3] A SMILES string is a series of letters that uniquely describes a molecular structure. SMILES strings can be transformed into vectors that are useful as predictors for chemical properties, or data the string can be used to look up or calculate other properties directly using libraries like RDKit.[4] RDKit is a multi-functional Python library for chemical studies. It can translate a SMILES string, a common English chemical name, or an IUPAC name into a set of molecular descriptors. Toropov's and our models are members of the class of models known as quantitative structure property relationships (QSPR) in which chemical properties are predicted directly from structural descriptors such as those calculated by RDKit. Several other researchers have previously developed QSPR relationships for Henry's law constants in water.[5] Many of these published models use classical group or bond contribution approaches[6] or radial basis function networks.[7] Others employ quantum chemical calculations to assist in generating data. Still others employ ML approaches to learn from large data sets.[8]

Expanding on concepts explored in previously published work, this study aims to expand QSPR capabilities in three major ways. First, the size of compiled data sets has increased dramatically with the release and updates of Sander's database, leading to the ability to study many members of many classes

[a] J. White, J. Graf, S. Haines, N. Sathitsuksanoh, R. Eric Berson, V. W. Jaeger
Chemical Engineering Department, University of Louisville, 216 Eastern Pkwy, Louisville, KY 40208, USA
E-mail: vance.jaeger@louisville.edu

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cplu.202400459

of organic chemicals. Second, our model works to predict k_H in a new solvent (ethanol) for which there are limited experimental data. Third, we take advantage of recent advances in ML training and deployment. These advantages are useful for cases such as distilled spirits in which the composition of the solvent greatly affects the volatility of solutes and for cases in which many chemical classes of solutes are present. Recent developments in artificial intelligence (AI), specifically within the realm of ML allow for fast, accurate transformations from input descriptors to output predictions. Many tools now exist to quickly test a variety of ML models and strategies. One such tool is the *PyCaret* library in Python.[9] PyCaret provides functions to quickly deploy and test a variety of common regression models, which in turn allows techniques to be compared side by side.

With these new AI/ML tools, it would seem a straightforward task to train models for predicting k_H in water and ethanol solvents. Unfortunately, to the best of the authors' knowledge there exists no large database of experimental k_H values for organic compounds in ethanol. Thus, other sources of training data are needed. For this task, we applied QM calculations. QM models have been shown to accurately calculate the solvation free energy of organic compounds in various solvents by using the PCM implicit solvation model.[10] However, accurate QM calculations are computational expensive. For a set of hundreds to thousands of compounds, the calculation cost is prohibitive for all but those who have access to large computing resources. It is advantageous to centrally train ML models with a large set of results from expensive calculations that can be packed and deployed with minimal computational cost to the end user. Furthermore, end users need not understand the complexities of QM calculations or ML models to quickly retrieve reasonable estimations of missing data on vapor-liquid equilibria.

In this study, we will use taste and odor compounds from bourbon whiskey to demonstrate the new model's applicability. Olfactory and gustatory sensations are intricately tied to vapor-liquid equilibrium and to interactions between molecules and sensing organs. The designation "bourbon" for whiskey is defined by U.S. law.[11] Bourbon whiskey is an American distilled spirit. The grain bill that is fermented to produce bourbon consists of at least 51% maize corn by weight. Other common grains include barley, wheat, and rye. Bourbon is aged in new charred oak barrels. It is common to age the whiskey for 2–5 years before bottling and sale. The charred oak imparts unique flavor and odor compounds to the product as it ages. Bourbon whiskeys contain a wide range of desirable and undesirable organic molecules arising not only from the years-long aging process[12] but also from congeners produced during the fermentation process.[13] Because distillation is a separation process based on vapor-liquid equilibrium, Henry's law estimates can be useful for the design and tuning of bourbon processes. During barrel aging, the equilibrium processes drive organic compounds out of the charred wood into the liquid and into the vapor. Ethanol, water, and odor molecules can slowly escape the barrel. Not only does this movement of chemicals out of the barrel affect the quality and quantity of the product, but it also affects the air quality inside and outside the buildings used for storage during aging. This highlights the importance of probing vapor-liquid equilibrium.

Poisson and Schieberle identified many of the most important flavor and odor compounds in bourbon whiskey.[14,15] Yang and coworkers further investigated a list of chemical compounds for identifying the extent to which a whiskey has been aged.[16] We will use the compounds identified by these previous researchers to test of our new ML/QM model on industry-relevant compounds.

Methods

Quantum Mechanical Calculations

QM calculations were performed in Gaussian 09 revision C.01.[17] A total of 6007 chemicals were selected from Sander's database[1] for producing QM data in both water and ethanol. Density functional theory (DFT) was used with the B3LYP functional.[18] B3LYP was selected because it has been widely tested and accepted for uses with small to medium sized organic molecules. A 6–311++G(d,p) basis set was used. This triple-zeta basis set with diffuse functions captures long-range interactions. The polarization functions (d,p) improve the electron distribution in covalent bonding. This selection of method and basis set balanced the computational cost and model accuracy for such a large set of molecules. The solvation free energy was calculated by employing a polarizable continuum model (PCM). The SMD variation of the integral equation formalism (IEFPCM)[19] allowed accurate calculation of solvation energies with appropriate treatment of the solvents' polarization effects. SMD better handles both electrostatic and non-electrostatic contributions to solvation compared to PCM. Solvation free energy (ΔG_{solv}) can be converted to k_H by applying Equation (2). R is the ideal gas constant and T is the temperature, which we set to the reference temperature of 298.15 K. We denote the natural logarithm as *log* rather than *ln*.

$$\boldsymbol{log}(\boldsymbol{k_H}) = -\frac{\boldsymbol{\Delta G_{solv}}}{\boldsymbol{RT}} \tag{2}$$

Solvation energetics are intricately tied with Henry's law because free energy is related to relative probabilities. As solvation energy decreases, molecules prefer to be in the liquid-solvated phase. As solvation energy increases, molecules prefer to be in the gas or vacuum phase. Thus, solvation energy can be compared for a given molecule in different solvents as a proxy for measuring relative solubilities.

Initial molecular configurations were generated as z-matrices from the RDKit library in Python. QM calculations were run for an initial set of 1477 chemicals in water and in ethanol. To estimate systematic error arising from inaccuracies in the input structures, 100 replicas were used with slightly varying geometries. These 100 replicas allowed for the possibility of multiple low-energy conformers. The standard deviation of the ΔG_{solv} across these 1477 molecules was on average 0.66 kcal/mol. This fluctuation is near the level of thermal noise (RT = 0.59 kcal/mol at 298.15 K) Thus, we judged that fewer replicas were sufficient for future sets of calculations. For all other molecules 10 replicas were used. In addition to the original 1477 molecules, a second set of 4530 molecules was sampled. The QM-derived ΔG_{solv} data for all tested molecules is included in a.csv file in the Supporting Information. The average standard deviation of ΔG_{solv} across these additional

4530 molecules was 0.60 kcal/mol, thus validating the use of fewer replicas.

The QM models we applied are appropriate for estimating ΔG_{solv} for systems in which there are no energetically accessible protonation states. For molecules in which these states are present, ΔG_{solv} is inaccurately calculated. Only neutral states of the molecules were assessed. Charged states were avoided because they tend to resist moving to the gas phase and because it is difficult to accurately predict the pK_a of the molecule *de novo*.[20] Thus, vapor-liquid equilibrium is difficult to assess via QM methods without further complicating considerations. Carboxylic acids, pyridines, and amines were three classes of chemicals that were common in the experimental database that were not described well by the QM models. Further the B3LYP functional does not describe halogen bonds well.[21] Halogen bonds increase the propensity of halogenated compounds to be solvated in water. We observed that as the number of halogen atoms in a molecule increases the QM model deviates further from experiments.

Data Preparation

Data for the ML models were curated through a pipeline consisting of the Python packages *PubChemPy* and *RDKit*, along with the dataset of Henry's law constants from Sander's database.[22] The PubChemPy package was used to query the PubChem database to obtain SMILES strings from the common names.[23] SMILES strings present a convenient way to reduce complex molecular structure into a string datatype, which is analogous to a word or sentence.[24] These SMILES strings were then passed through RDKit to obtain structural and global information about each of the molecules.

Several key features are presented in Table 1, and a full list of descriptors is available in the Supporting Information. The data gathered from RDKit fit into two categories, namely (a) quantities of molecular groups, and (b) macroscopic molecular properties like molecular weight and charge characteristics. Molecular groups were selected based on the available fragments identifiable within RDKit. Macroscopic properties that were likely to affect solvation were also selected from RDKit. These data were used to create the input vector to be applied with a set of eighteen ML regressors. Because the model includes extensive data about chemical functional groups, the strategy we have employed resembles a hybrid group contribution approach.

Ground-truth Henry's law constants were taken from Sander's database and used in supervised learning and to test network performance. The database contains data on over 10,000 species, but we selected a subset of the data. Molecules that had two or fewer data points were excluded. This, combined with availability of SMILES strings from PubChem, reduced the number of molecules to 5,690 for ML models. Because this dataset included data for Henry's law constants exclusively in water, a small dataset was collected from existing literature on a few example molecules in ethanol.[25] This test dataset was severely limited by the available literature. After the data were compiled, the data were min-max scaled across each feature dimension prior to use. PyCaret was used to rapidly prototype 18 traditional machine learning models. Models were compared based on a variety of accuracy measures including mean absolute error and R^2. The top four of these models were selected for further fine-tuning and analysis. These top models were extra trees, ridge, Bayesian ridge, and huber regressors. These models were further tested and fine-tuned via a random split of the dataset into training (90%) and testing (10%) sets with a specific random seed for reproducibility.

Table 1. Tabulated summary of included features for model prediction.

Vector Index	Features
0–85	Data on quantity of 86 molecular fragments output from RDKit Descriptors (e.g. # of alcohols, # of amines, # of aldehydes, etc.)
86	molecular weight
87	heavy molecular weight
88	minimum absolute partial charge
89	maximum absolute partial charge
90	number of valence electrons

Results and Discussion

Validating QM Calculations

Because there are extensive data for k_H in water but not ethanol, it is important to first validate the use of QM models for predicting k_H in water. To validate the QM calculations, QM-derived ΔG_{solv} values were compared to k_H. As shown by Equation (2), the $\log(k_H)$ should equal $-\Delta G_{solv}/RT$. Rather than an exact match between these values, our results in Figure 1 indicate that there is a linear relationship of as in Equation (3).

$$\log(k_H) = -0.59\frac{\Delta G_{solv-water-QM}}{RT} + 2.23 \quad (3)$$

The linear fit has a Pearson correlation of value of $R^2 = 0.71$, indicating a reasonable linear fit. The mean absolute error among the linear fit predictions is 1.79 in the log scale, which is comparable to that of the average experimental range of 1.93

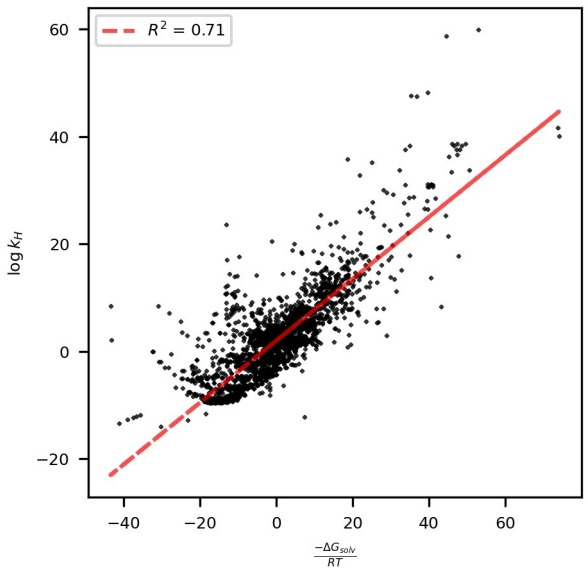


Figure 1. Linear relationship between Sander's compilation of experimental and modelled Henry's law constant (k_H) in water and solvation free energy ($\Delta G_{solv\text{-}water}$) in water from quantum mechanical calculations. R and T are the ideal gas constant and temperature respectively. Error bars are excluded for clarity but can be found in the Supporting Information.

in the log scale. Full data on the experimental error ranges are available in the Supporting Information within.csv files. The fact that there is a shift in the slope and intercept is not surprising. Previous studies have shown that systematic errors in QM methods and basis sets lead to shifts in energetics. However, others have found similar linear relationships hold for different classes of chemicals.[26] With access to QM and the expertise to run such calculations, Equation (3) can be used to correct systematic errors for solvation energies using the same method, basis set, and implicit solvent as us (B3LYP/6-311++G(d,p)) with SMD implicit solvent).

Comparing QM Models for Water and Ethanol Solvents

QM calculations supplement our models to allow calculations of k_H in ethanol by creating a proxy versus water. The solvation free energy relates directly to a compound's propensity to prefer a particular solvent. The energetics are a function of the various intermolecular forces within the solvent and between the solvent and the solute. The major enthalpic contributions in the solution phase come from hydrogen bonding and dipoles. In this way, ethanol and water present two, similar solvent environments. Therefore, it is likely that the solvation energy of a molecule within them will be related. Figure 2 indicates that there is a strong linear relationship between ΔG_{solv} of a molecule in water and ΔG_{solv} in ethanol. The relationship is provided in Equation (4).

$$G_{solv-ethanol-QM} = 0.79\ \Delta G_{solv-water-QM} - 3.10 \quad (4)$$

The linear fit between solvation free energies in water and ethanol based on QM calculations has a Pearson correlation of $R^2 = 0.97$. The observed quality of fit is likely attributable to the chemical similarities between water and ethanol, as well as the use of similar QM calculation methods. The tighter fit observed in Figure 2 compared to Figure 1 reflects the difficulty in comparing experimental data from multiple sources and methods to data generated from a single source with a single QM method. The strong correlation and slope of 0.79 suggest that ethanol and water solvate molecules in a similar way with ethanol being slightly effective as a solvent. Water tends to solvate hydrophilic molecules more favorably than ethanol, while ethanol may be slightly better at solvating molecules with nonpolar regions due to ethanol's amphiphilic nature. For molecules with highly negative solvation energies in water, which are typically hydrophilic, solvation favors the liquid phase over the gas phase. Conversely, hydrophobic molecules, which have higher solvation energies in water, are less stabilized in the liquid phase. The slope indicates that ethanol provides weaker overall solvation compared to water for polar molecules, due to ethanol's lower polarity and fewer hydrogen bonds. However, the intercept of −3.10 kcal/mol suggests that ethanol can solvate nonpolar regions of molecules more effectively than water, likely due to dispersion interactions from ethanol's hydrophobic ethyl group.

Datapoints whose standard deviation in solvation energy across confirmations was greater than 10 kcal/mol were excluded from this correlation. Visual inspection of the final molecular structures from QM showed that molecules with large standard deviations across 10–100 replicas typically underwent bond rearrangement to form an isomer during the calculation. By applying Equation (4), one can translate between ΔG_{solv} in water and ΔG_{solv} in ethanol. Further, by assuming the relationship in Equation (3) holds in both solvents, it is straightforward to translate between experimental, ML, or theoretical k_H values in water and in ethanol.

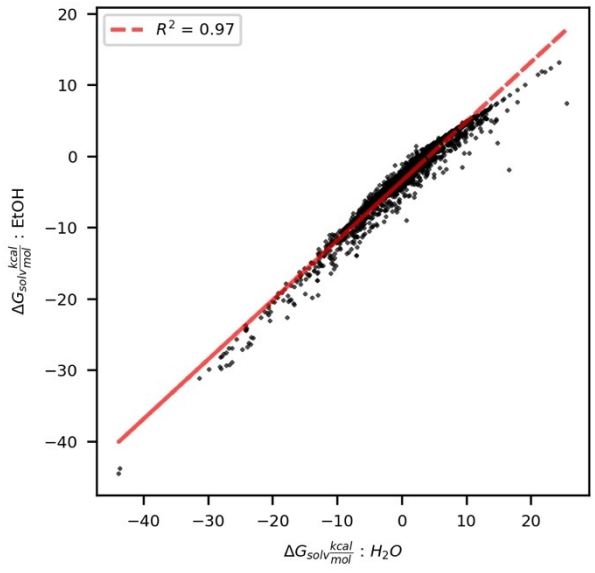


Figure 2. Linear relationship between solvation free energy (ΔG_{solv}) from quantum mechanical calculations in implicit water and ethanol solvents.

ML Model for Predicting k_H in Water

Structural descriptors provided by RDKit along with macroscopic molecular properties were used to train a set of 18 standard regression models using the PyCaret library's compare-models functionality. These regression strategies include: ridge, Bayesian ridge, Huber, light gradient boosting machine, extra trees, gradient boosting, random forest, passive aggressive, orthogonal matching pursuit, k nearest neighbors, decision tree, adaboost, elastic net, lasso, lasso least angle, dummy, linear, and least angle. The top performing regression models were found to be, extra trees, ridge, Bayesian ridge, and Huber regression. Each modeling technique was refined and refitted using 90% of the total data, with the remaining 10% reserved for testing. The refinement process involved 10 iterations of 10-fold cross-validation, during which the R^2 values were monitored for data points excluded in each fold. The space of possible hyperparameters was traversed using random grid search. After the models were fit, each was tested on the remaining 10% of data. Figure 3 demonstrates the accuracy of these models in predicting on the testing set. The best performing model out of these four was the extra trees regression, with $R^2 = 0.98$, and a mean absolute error of 1.3 in

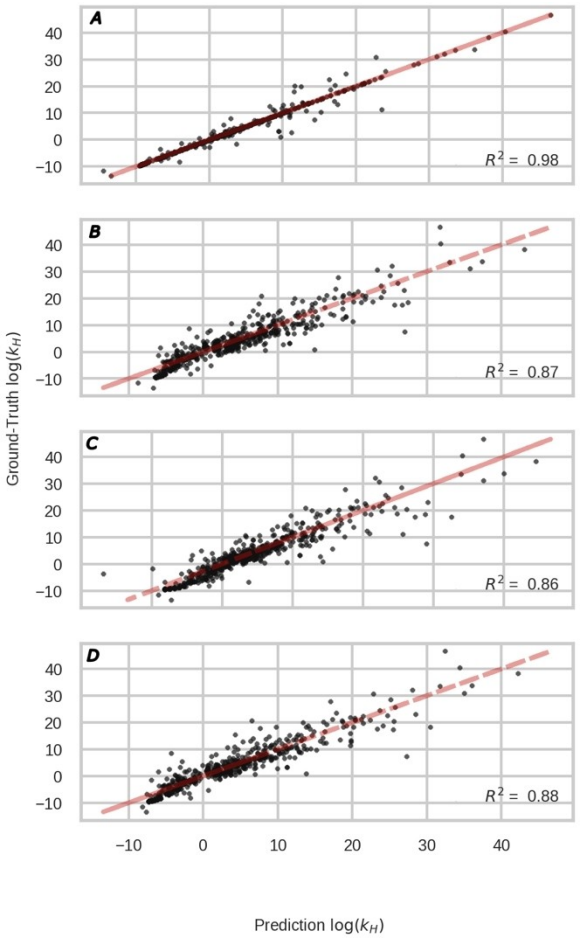


Figure 3. Machine learning models for predicting Henry's law constant (k_H) directly from structural descriptors. (A–D) Extra trees regression, ridge regression, Bayesian ridge, and Huber regressor. These data represent 10% of the total set (569 molecules). Red line represents a perfect model.

log space. Extra trees regression presents a type of tree-based ensemble model, akin to random forest. Like others, the tunable parameters relate to the size of the tree, width, depth, and number of leaves. The final model was an ensemble of 100 trees with an average node count of 6544 nodes. Further information on the four modeling techniques explored can be found in the Supporting Information (Table S1). The model error is similar to the average range of 1.93 across experiments in the database.

To enhance model interpretation, feature importance for the Extra Trees model was analyzed, as detailed in Table 2. Feature importance was determined by calculating the mean cumulative reduction in impurity across the ensemble of trees.[27] This approach reflects the increase in predictive certainty attributable to each feature, as measured in the nodes where the feature appears. This method is a standard practice for assessing feature importance in tree-based classifiers and regressors. The two most important features were found to be the maximum and minimum partial charge magnitude. This finding aligns with the understanding that localized charge characteristic increases the energy required for a molecule to transfer from the solvated liquid to the gas phase. Two additional important features are the quantities of both tertiary and non-tertiary aliphatic hydroxyl groups. This is significant as these groups participate in hydrogen bonding, and therefore increase the propensity for the molecule to remain in the solvated phase.

Table 2. Tabulated relative feature importance for every feature with greater than 0.020 importance, in descending order.

Feature (N=Number of)	Relative Importance Value
maximum partial charge magnitude	0.245
minimum partial charge magnitude	0.127
N hydroxyls	0.057
N benzenes	0.056
N carbonyls	0.050
N non-tertiary hydroxyls	0.048
N primary amines	0.045
N secondary amines	0.042
N aromatic nitrogens	0.025
N tertiary amines	0.025
N halogens	0.020

Predicting k_H in Ethanol

To perform the predictions in an ethanol solution, the extra trees model was then paired with a linear translation layer formed via the combining of Equations (2) and (4) as described in Equation (5). This combination is made under the assumption that the relationship between solvation energy in water and ethanol found from the QM simulations holds for experimental values. This combination is reduced taking R to be 0.001986 kcal/mol-K and T to be 298.15 K. The full workflow, from molecule name to predicted kH, is outlined in Figure 4.

$$\boldsymbol{log}(\boldsymbol{k_H})_{\boldsymbol{ethanol}} = 0.79\boldsymbol{log}(\boldsymbol{k_H})_{\boldsymbol{water}} + 5.25 \quad (5)$$

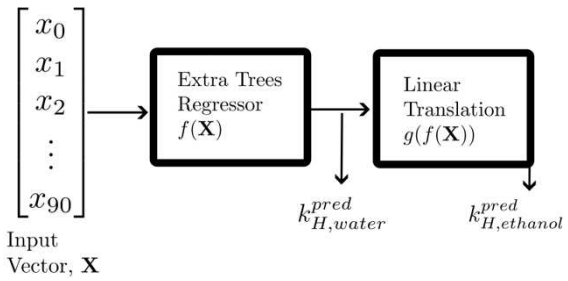


Figure 4. Basic structure of prediction flow, including the input vector, regression, and linear translation layers.

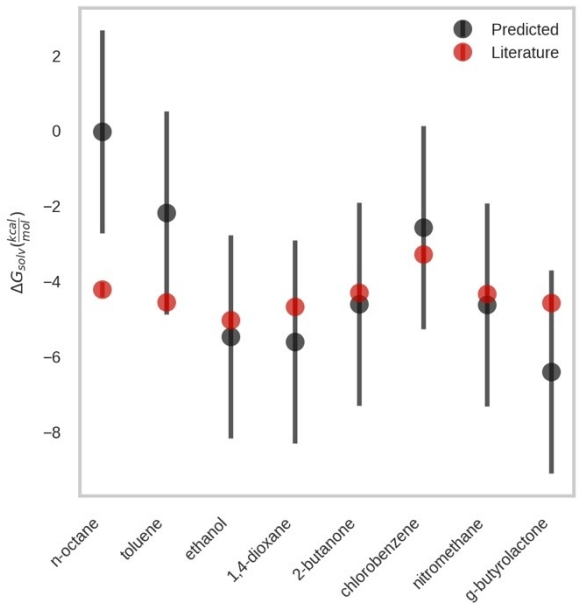


Figure 5. Predicted and literature ΔG_{solv} values to test model's inference capability. Error bars on literature data are based on documentation in the database. Error bars on predictions are estimated based on mean absolute model error combined with mean absolute error in the translation from water to ethanol.

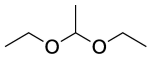
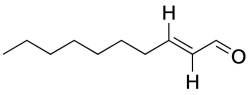
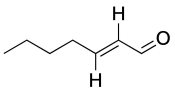
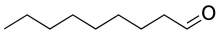
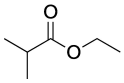
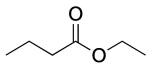
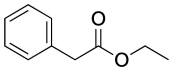
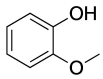
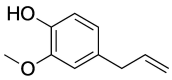
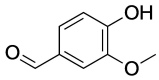
Validation of Ethanol Model with a Small Data Set

To validate the complete model, several experimental data points were found in the literature.[25] Each molecule was subsequently passed through the pipeline as outlined in Figure 4, resulting in a prediction of the Henry's law constants in ethanol. For comparison, the values were converted to solvation energies rather than fully into Henry's law constants, as more relevant data is available for the former. Predictions are compared to literature data in Figure 5. Error ranges for predicted values are calculated based on the error introduced through the linear translation layer. For the experimental values, the error is described in the database. Most predictions fit reasonably well with literature values.

Example Use Case with Bourbon Compounds

To provide an example of the use case for our model, an existing set of odorants relevant to distilled spirits was tested. Molecule names listed in Table 3 were passed through the prediction pipeline. The overall computation time for the pipeline was less than one minute. Thus, the technique

Table 3. Tabulated predicted values for k_H for bourbon related odorants.[1,14,28] Experimental ranges represent the full range of the values reported by Sander. Exp. denotes experimental or model values reported by Sanders and Pred. denotes predicted values from our ML pipeline.

Name	Structure	Pred. logk_H (Wat.)	Exp. logk_H (Wat.)	Exp. range logk_H (Wat.)	UNIF logk_H (Wat.)	Pred. logk_H (EtOH)	UNIF logk_H (EtOH)
1,1-diethoxyethane (acetal)		2.17	2.15	1.29	0.0729	6.97	4.67
(E)-2-decenal		2.24	-	-	1.85	7.02	10.9
(E)-2-heptenal		1.65	1.62	0.2	1.49	6.55	7.73
1-nonanal		0.879	0.938	3.99	0.87	5.95	9.32
ethyl 2-methylpropanoate (ethyl isobutyrate)		0.544	0.590	0.594	0.00586	5.68	4.99
ethyl butanoate (ethyl butyrate)		0.823	2.48	5.68	−0.395	5.91	5.21
ethyl 2-phenylacetate		5.25	-	-	1.30	9.41	9.69
2-methoxyphenol (guaiacol)		8.23	-	-	9.13	11.8	12.7
4-allyl-2-methoxyphenol (eugenol)		6.78	-	-	14.6	10.69	12.7
4-hydroxy-3-methoxy-benzaldehyde (vanillin)		13.9	14.0	6.11	14.6	16.3	17.5

produces rapid estimates of vapor-liquid equilibrium characteristics without the need for additional ML model tuning, QM calculations, or experimental data. The tested compounds directly relate to the scent and flavor-profiles of Bourbon whiskey. The model predictions were compared to experimental data where available in the case of water and against UNIFAC predicted results in the case of ethanol. As can be seen, the model predicted values align reasonably well with water, with a mean absolute error of 0.31 in log space. In ethanol, the model predicts values reasonably like UNIFAC, with a mean absolute error of 1.65 in log space. Overall, the model presents a faster and easier means of estimating Henry's law constants as compared to experiments, QM calculations, and UNIFAC models.

Determining k_H in ethanol for components where experimental data is lacking by leveraging a validated correlation with k_H in water values extends the applicability of the ML model beyond which experimental validation exists. Having k_H data, whether validated directly or indirectly, allows for insights into the solubility and partitioning behavior of these components in ethanol for which reliable methods are currently unavailable. More generally, this methodology highlights the versatility of QM and ML in predicting thermodynamic properties across various solvent systems where experimental data may be limited or nonexistent.

Conclusions

In this study, a new ML model was developed for the prediction of Henry's law constants in water and ethanol. This new model allows researchers to quickly estimate k_H for a wide range of organic chemicals. By supplementing experimental data with QM calculations, new solvent systems like ethanol can be explored using these models. The new model allows for rapid searches over large sets of molecules related to the distilled spirits industry. The discovered strong relationship between water and ethanol solvation free energies may also point toward the possibility of developing translation layers between QM simulations, allowing for one simulation to generate a disproportionate amount of data.

There are a few important limitations to note. First, molecules that contain a net charge near neutral pHs are not well represented in QM calculations. For this reason, we suggest not using the model for carboxylic acids, pyridines, and amines in ethanol. However, the limitations listed are not exhaustive, and users should use their own intuition and knowledge about the pK_a(s) of any test molecule. Second, molecules that contain halogens are not represented well by the QM methods used in generating the models. We suggest not using halogenated compounds in any ethanol predictions. For food and beverage applications, we expect that halogenated compounds are of limited interest. Because the ML model for water solvent is not based upon QM calculations, the ML model should hold for any of the classes (charged or halogenated) excluded from the ethanol models. Third, this model predicts behavior at a reference temperature of 298.15 K. We expect trends to hold for other temperatures near 298.15 K, but this is not guaranteed. This model is also limited in its ability to make predictions on noble gases, as the feature set is not rich enough to differentiate this class of compounds. Noble gases, however, comprise much of the available experimental data for k_H in non-aqueous solvents, which limits the need for a predictive model. Enantiomers are also not differentiated by the model, but it is unlikely that two enantiomers of the same molecule have significantly different vapor-liquid equilibrium behavior. Finally, these models hold for compounds at low concentrations dissolved in pure water or in pure ethanol. As the concentration of a solute increases, deviation from Henry's law is expected.

In future work, we plan to revise our model to include adjustments for identified weaknesses and to demonstrate the applicability of the model to additional systems within the fields of food science and environmental science. Shortcomings in modeling charged compounds can be ameliorated by considering both charged and uncharged states in the QM calculations, thereby allowing titration to be studied. Other QM functionals that better model halogen bonding can be used in lieu of B3LYP. Some data included in the Henry's law database describes the temperature dependence of k_H. Inclusion of this variable would improve predictions at temperatures away from the reference temperature of 298.15 K. However, there is a limited amount of experimental data to train new models for ethanol. The methods we demonstrate here could be applied to any solvent that is well-represented in QM calculations by PCM models. It will be important for future work to be able to provide predictions not only for pure solvents but also for mixtures. The accuracy of mixing rules should be explored. Future work will also include exploring an expanded set of noteworthy features, to extend the breadth of chemicals on which the model is effective. This could include information on hydrophobicity, pK_a, isomerization, and many more. Utilizing a more robust data-driven method, such as one of the presented ML frameworks, for the conversion of solvation energy from water to ethanol should also be explored.

Supporting Information Summary

Data generated by QM calculations for both water and ethanol are provided in the Supporting Information as .csv files. This catalog of data may be useful for future researchers who would like to explore the solvation of small molecules in water and ethanol. Additional interesting ML models for the prediction of phase equilibria could be generated using these data.

The ML models developed in this paper are included as Python scripts and pickle files in the Supporting Information including a short how-to guide on applying the models to new chemicals. Included along with this manuscript are .csv datafiles with the relevant data used in this study, along with a list of all the features used in the input token. There is also a folder with a usable python command line application version of the tool described in the manuscript with brief usage instructions.

Acknowledgements

The authors acknowledge computational resources provided by the Cardinal Research Cluster which is supported by the University of Louisville's Office of Research and Innovation.

Conflict of Interests

The authors declare no conflict of interest.

Data Availability

In addition to being available in the Supporting Information, the raw data used in this study is available on GitHub. https://github.com/jwthe3rd/QSPRPaperData.

Keywords: Quantitative Structure Property Relationships · Computational Chemistry · Food Science · Machine Learning · Henry's Law

[1] R. Sander, *Atmos. Chem. Phys.* **2023**, *23*, 10901–12440.
[2] J. T. Dickman, Masters Thesis, University of Southampton **2020**.
[3] A. A. Toropov, A. P. Toropova, A. Roncaglioni, E. Benfenati, D. Leszczynska, J. Leszczynski, *Molecules* **2023**, *28*, 7231.
[4] G. Landrum, RDKit, Program for Open-source cheminformatics, **2023**.
[5] a) A. R. Katritzky, L. Mu, M. Karelson, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1162–1168; b) M. H. Abraham, J. Andonian-Haftvan, G. S. Whiting, A. Leo, R. S. Taft, *J. Chem. Soc., Perkin Trans. 2* **1994**, *8*, 1777–1791.
[6] a) S.-T. Lin, S. I. Sandler, *Chem. Eng. Sci.* **2002**, *57*, 2727–2733; b) W. M. Meylan, P. H. Howard, *Environ. Toxicol. Chem.* **1991**, *10*, 1283–1293.
[7] a) X. Yao, M. Liu, X. Zhang, Z. Hu, B. Fan, *Anal. Chim. Acta* **2002**, *462*, 101–117; b) H. Modarresi, H. Modarress, J. C. Dearden, *Chemosphere* **2007**, *66*, 2067–2076.
[8] a) D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas, F. Giralt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 85–112; b) N. J. English, D. G. Carroll, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1150–1161; c) R. Petersen, A. Fredenslund, P. Rasmussen, *Comput. Chem. Eng.* **1994**, *18*, S63–S67.
[9] M. Ali, PyCaret, An Open Source, Low-code Machine Learning Library in Python, **2020**.
[10] a) J. Ho, M. Z. Ertem, *J. Phys. Chem. B* **2016**, *120*, 1319–1329; b) S.-C. Liu, X.-R. Zhu, D.-Y. Liu, D.-C. Fang, *Phys. Chem. Chem. Phys.* **2023**, *25*, 913–931.
[11] Title 27, Chapter 1, Subchapter A, Part 5, Code of Federal Regulations, United States.
[12] a) T. Tarko, F. Krankowski, A. Duda-Chodak, *Molecules* **2023**, *28*, 620 – 643; b) J. Gollihue, V. G. Pook, S. DeBolt, *J. Inst. Brew.* **2021**, *127*, 210–223; c) M. Luo, D. Cui, J. Li, P. Zhou, C. Duan, Y. Lan, G. Wu, *Foods* **2023**, *12*, 4266–4281.
[13] a) T. J. Kelly, C. O'Connor, K. N. Kilcawley, *Beverages* **2023**, *9*, 64; b) M. Stockwell, I. Goodall, D. Uhrín, *Anal. Sci. Adv.* **2020**, *1*, 132–140; c) D. González-Arjona, V. González-Gallero, F. Pablos, A. Gustavo González, *Anal. Chim. Acta* **1999**, *381*, 257–264.
[14] L. Poisson, P. Schieberle, *J. Agric. Food. Chem.* **2008**, *56*, 5820–5826.
[15] L. Poisson, P. Schieberle, *J. Agric. Food. Chem.* **2008**, *56*, 5813–5819.
[16] K. Yang, A. Somogyi, C. Thomas, H. Zhang, Z. Cheng, S. Xu, C. Miller, D. Spivey, C. Blake, C. Smith, D. Dafoe, N. D. Danielson, M. W. Crowder, *Food Anal. Methods* **2020**, *13*, 2301–2311.
[17] M. J. Frisch, G. Trucks, H. B. Schlegel, G. E. Scuseria, Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, Gaussian 09 rev. 1, Program for performing MD Simulations, **2009**.
[18] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.
[19] A. V. Marenich, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. B* **2009**, *113*, 6378–6396.
[20] a) S. Pezzola, M. Venanzi, P. Galloni, V. Conte, F. Sabuzi, *Molecules* **2024**, *29*, 1255; b) R. Lawler, Y.-H. Liu, N. Majaya, O. Allam, H. Ju, J. Y. Kim, S. S. Jang, *J. Phys. Chem. A* **2021**, *125*, 8712–8722.
[21] A. Otero-de-la-Roza, E. R., Johnson, G. A. DiLabio, *J. Chem. Theory Comput.* **2014**, *10*, 5436–5447.
[22] M. Swain, PubChemPy, Program for retrieving chemical information, **2024**.
[23] S. Kim, J. Chen, T. Cheng, *Nucleic Acids*, **2023**, *51*, D1373–D1380 .
[24] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
[25] P. Winget, D. M. Dolney, D. J. Giesen, C. J. Cramer, D. G. Truhlar; **1999**; Minnesota solvent descriptor database; Minneapolis, MN: Department of Chemistry and Supercomputer Institute.
[26] X. Zhang, Y. Zeng, *Fluid Phase Equilib.* **2014**, *376*, 234–238.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
[28] A. Fredenslund, R. L. Jones, J. M. Prausnitz, *AIChE J.* **1975**, *21*, 1086–1099.

Manuscript received: July 3, 2024
Revised manuscript received: September 16, 2024
Accepted manuscript online: September 20, 2024
Version of record online: November 1, 2024