QSPR Molecular Approach for Estimating Henry's Law Constants of Pure Compounds in Water at Ambient Conditions

ARTICLE *in* INDUSTRIAL & ENGINEERING CHEMISTRY RESEARCH · MARCH 2012

Impact Factor: 2.59 · DOI: 10.1021/ie202646u

I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

QSPR Molecular Approach for Estimating Henry's Law Constants of Pure Compounds in Water at Ambient Conditions

Farhad Gharagheizi,*[,†] Poorandokht Ilani-Kashkouli,[‡] Seyyed Alireza Mirkhani,[†] Nasrin Farahani,[§] and Amir H. Mohammadi*[,∥,⊥]

[†]Department of Chemical Engineering and [§]Department of Chemistry, Islamic Azad University, Buinzahra Branch, Buinzahra, Iran

[‡]Department of Chemical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[∥]MINES ParisTech, CEP/TEP - Centre Énergétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France

[⊥]Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa

Supporting Information

ABSTRACT: In this article, we present a comprehensive quantitative structure−property relationship (QSPR) to estimate the Henry's law constant (H) of pure compounds in water at ambient conditions. This relationship is a multilinear equation containing eight chemical-structure-based parameters. The parameters were selected by the genetic algorithm multivariate linear regression (GA-MLR) method using more than 3000 molecular descriptors. The squared correlation coefficient of the model (R^2) over 1954 pure compounds is equal to 0.983 (logarithmic-based data). Therefore, the model is comprehensive and accurate enough to be used to predict the Henry's law constants of various compounds in water.

1. INTRODUCTION

The Henry's law constant of a pure compound in water, or the air−water partition coefficient (H), is defined as the ratio of the partial pressure of a chemical compound in air to the concentration of that compound in water at a given temperature. This parameter is used to describe the tendency of a chemical compound to volatilize from the liquid water surface into the atmosphere. Hence, chemical compounds with higher H values are more likely to volatilize from aqueous solutions.[1,2]

Accurate knowledge of Henry's constants enables scientists to study the movement of chemical compounds inside and outside aquatic ecosystems such as ponds, lakes, and oceans.[1,2] This is mainly significant in the environmental sciences. The H value is a good indicator of a chemical compound's volatility, which demonstrates the tendency of the compound to leave an aqueous solution, as presented by Lyman et al.:[3]

- For $\log(H) < -7$, the chemical compound is less volatile than water, and its concentration will increase as water evaporates; it is essentially nonvolatile.
- For $-7 < \log(H) < -5$, the chemical compound slowly volatilizes,at a rate controlled by slow molecular diffusion through air.
- For $-5 < \log(H) < -3$, volatilization begins to become a significant transfer mechanism; this range contains most polycyclic aromatic hydrocarbons and halogenated aromatics.
- For $-3 < \log(H)$, chemical compounds are released in significant quantities; resistance from the water film is the rate-controlling process.

In the above statements the unit of H is atm·m^3/mol. Accurate measurements of H may be difficult and expensive partly because of some practical problems such as the adsorption of small amounts of solute on the wall of the apparatus and analytical detection limits of low concentrations of highly hydrophobic compounds.[4,5] In addition, based on the literature, experimental values of H have been reported for a relatively small number of pure compounds. Therefore, the development of computational methods for the representation/prediction of this parameter for a wide range of pure compounds is essential.[6]

Many predictive tools have been proposed for the prediction of H and can be categorized into two main types according to their parameters: The first type contains correlations using physical properties such as vapor pressure and aqueous solubility to correlate the H values. The methods proposed by Mackay et al.[4] and Gharagheizi et al.[7] are of this class. These correlations, however, have some shortcomings. In particular, the accuracy of the predictions of these correlations depends directly on the accuracy of the evaluated physical properties. Furthermore, if only one of the required properties is unavailable, no calculation can be performed to predict H.

The second type consists of quantitative structure−property relationships (QSPR). The latter employs parameters called "molecular descriptors" to describe chemical structure. Developing predictive tools using these parameters is the main idea that is pursued in QSPR. The most well-known models in this category are those presented by Hine and Mookerjee,[6] Meylan and Howard,[8,9] Abraham et al.,[10] Katritzky et al.,[11] Dearden et al.,[12,13] English and Carrol,[14] Yao et al.,[15] Bernazzani et al.,[16,17] Lin and Sandler,[18] Yafe et al.,[19] Modarresi et al.,[20] and Gharagheizi et al.[21] The most important disadvantage of the majority of these methods is their complex

Received: November 16, 2011
Revised: February 19, 2012
Accepted: February 29, 2012
Published: February 29, 2012

dx.doi.org/10.1021/ie202646u | *Ind. Eng. Chem. Res.* 2012, 51, 4764−4767

computation procedure for determining parameters from chemical structures. However, the second type of tool generally shows more reliable results than the first type.[7,21]

In this work, the QSPR technique is employed to develop a model for predicting the Henry's constants of various pure compounds in water at ambient conditions. To develop this model, an extensive database containing 1954 pure compounds was used. The comprehensiveness of the presented model is demonstrated herein.

2. MATERIALS AND METHOD

2.1. Data Set. The comprehensiveness of a molecular-based model depends directly on the comprehensiveness of the data set of compounds used in its development. This characteristic includes both the diversity in the chemical families used and the number of compounds available in the data set. Our literature survey shows that one of the most comprehensive data sets available for H parameter values is the compilation provided by Yaws.[22] The Yaws compilation is one of the most reliable sources of experimental data gathered from more than 1000 high-quality references. The H values of 1954 pure compounds were extracted from this database and used as the main data set in this work. It should be noted that the H values were compiled in units of atm·m^3·mol^{-1} (mole basis) and are presented as decimal logarithms of H at ambient conditions. The applied data set is presented in the Supporting Information.

2.2. Determination of Molecular Descriptors. In this step, the molecular structures of all 1954 pure compounds were drawn using Hyperchem software[23] and optimized using the MM+ molecular mechanics force field. Because the values of some types of molecular descriptors depend on bond lengths, bond angles, and other structural characteristics, the real values for these parameters were required. Therefore, after the optimization of molecular structures, the molecular descriptors were calculated using Dragon software.[24] The latter can calculate more than 3000 molecular descriptors for each molecule. More information about the types of molecular descriptors and the procedure for computing the descriptors can be found elsewhere.[24]

2.3. GA-MLR Calculations. To develop a QSPR model for prediction of H, one desires a linear equation that can predict H values with the smallest number of variables and the highest accuracy. In other words, the problem is to obtain an optimum subset of variables (most statistically effective molecular descriptors for H) from all available variables (all molecular descriptors) to estimate H values with a minimum deviation from the available experimental data.

A generally accepted method for solving the aforementioned problem is genetic algorithm multivariate linear regression (GA-MLR) approach. In this method, a genetic algorithm is used to select the best subset of variables with respect to an objective function. Application of this algorithm for variable subset selection was first presented by Leardi et al.[25]

In this work, the GA-MLR algorithm with the RQK objective function presented by Todeschini and Consonni and Todeschini et al.[26,27] was used for the variable subset selection. The fitness function was defined as the leave-one-out cross validation coefficient (Q_{LOO}^2) subject to four constraints that must be fulfilled to avoid chance correlations and lack of predictive power. This method has been extensively applied in previous works by Gharagheizi and co-workers.[28−36]

According to the GA-MLR technique, the data set should be divided into two new collections. The first is assigned to training purposes, and the second is assigned to testing purposes. The best model is found by the training set, and the predictive power of the obtained model is later evaluated using the test set. In this work, 80% of the database was used for the training set, and 20% was used for the test set. The selection for the two sets was done randomly.[37]

The input parameters of our program were the pool of molecular descriptors, the H values (of the training set), and the desired number of molecular descriptors in our final model. To obtain the best multivariate linear equation, all molecular descriptors were introduced into the program, and the minimum number of possible variables was tested at the starting point. Consequently, the program was first run to obtain the best multivariate linear model using one variable. In the next steps, the number of desired variables was increased to two, three, four, and so on, and all computation steps were repeated.

The most reliable model is the one whose accuracy is not considerably affected by an increase in the number of its parameters.

3. RESULTS AND DISCUSSION

The most accurate multivariate linear equation was obtained by the procedure presented in the preceding section. The identified multivariate linear model had eight parameters as follows

$$\begin{aligned} H = {} & -1.8117(\pm 0.0543) - 1.2073(\pm 0.0182)\mathrm{ESpm01d} \\ & + 0.2234(\pm 0.0053)\mathrm{J3D} - 1.0379(\pm 0.0397)\mathrm{HOMA} \\ & + 0.5384(\pm 0.0319)\mathrm{nR{=}Ct} - 1.7409(\pm 0.0522)\mathrm{nRNHR} \\ & - 1.3957(\pm 0.0449)\mathrm{nHDon} - 0.9146(\pm 0.0486)\mathrm{Hy} \\ & - 1.7840(\pm 0.0250)\mathrm{B01[C\text{-}O]} \end{aligned} \quad (1)$$

$$n_{\mathrm{training}} = 1564, \quad n_{\mathrm{test}} = 390, \quad R_{\mathrm{training}}^2 = 0.9827, \quad R_{\mathrm{test}}^2 = 0.9832$$

$$Q_{\mathrm{LOO}}^2 = 0.9824, \quad Q_{\mathrm{BOOT}}^2 = 0.9822, \quad Q_{\mathrm{EXT}}^2 = 0.9828$$

$$s = 0.285, \quad a = -0.013, \quad F = 11018.781$$

RQK function parameters

$$\Delta K = 0.077, \quad \Delta Q = 0.000, \quad R^{\mathrm{P}} = 0.005, \quad R^{\mathrm{N}} = 0.000$$

where, as mentioned previously, H is in atm·m^3·mol^{-1} (mole basis).

In eq 1, ESpm01d is the spectral first moment from the edge adjacency matrix weighted by dipole moments (a measure of polarity).

J3D is the three-dimensional Balaban index. It is computed based on the geometric vertex distance degree, which is also a measure of branching. The branching is a direct result of an increase in the number of vertices in a molecule. Therefore, when branching increases, the number of vertices increases, and as a consequence, the three-dimensional Balaban index increases. Equation 1 shows that the H value increases with increasing branching in a molecule.

dx.doi.org/10.1021/ie202646u | Ind. Eng. Chem. Res. 2012, 51, 4764−4767

HOMA is the harmonic oscillator model of aromaticity index. The index is based on the degree of alternation of single and double bonds, measuring the bond length deviation from the optimal length attributed to the typical aromatic state. It is a measure of aromaticity, and when it increases, the tendency toward water increases, and therefore, the H value decreases.

nR=Ct is the number of aliphatic tertiary carbon atoms (sp^2), and nRNHR is the number of secondary amines (aliphatic amines). Both of these descriptors are group-count descriptors. The first is a measure of branching, thus it functions analogously to J3D. The second is a measure of weak hydrogen bonding. Its increase causes an increase in the tendency toward water, thus decreasing H.

nHDon is the number of donor atoms for H-bonds (N and O). It is one of the most well-known descriptors to show the effects of H-bonding. H-bonds clearly cause an increasing tendency toward water, therefore when this parameter increases, the H value decreases.

Hy is a hydrophilic factor. When it decreases, the tendency toward water increases. Consequently, the H value decreases.

B01[C-O] is a fingerprint-type descriptor. When it increases, the tendency toward water increases, thus the H value decreases.[26]

$n_{training}$ and n_{test} are the numbers of compounds in the training set and in the test set, respectively.

To better evaluate the validity of the model, the bootstrap technique, y-scrambling, and external validation techniques were used.[26] Bootstrapping was repeated 5000 times, and y-scrambling was repeated 300 times. As can be seen, the differences between Q_{LOO}^2, Q_{BOOT}^2, Q_{EXT}^2, and $R_{training}^2$ show that the obtained model has good predictive power. Moreover, the intercept of the y-scrambling technique has a low value ($a = -0.013$) that reveals the validity of the model. (The y-scrambling, bootstrapping, and external validation techniques were presented in detail by Todeschini and Consonni.[26])

All of the validation techniques used show that the obtained model is a valid model and can be reliably used to predict the H values of pure compounds.

The values of H predicted using eq 1 in comparison with those values reported by Yaws[22] are presented in Figure 1.

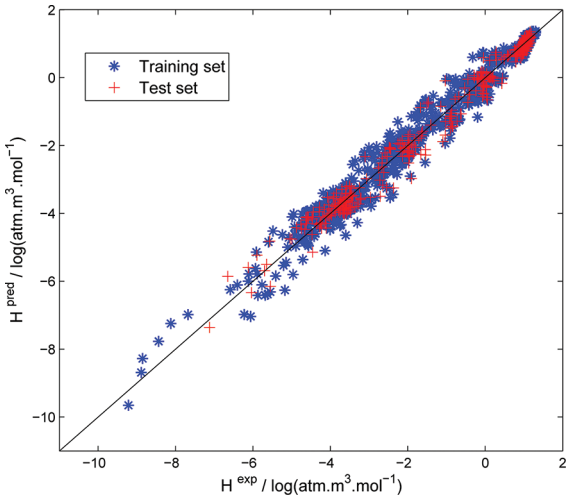

Figure 1. Comparison between H values predicted using eq 1 (superscript pred) and H values reported by Yaws[22] (superscript exp).

Moreover, the values of the descriptors and statuses of all of the pure compounds (training set or test set) are presented in the Supporting Information.

Some important points should be considered in comparing this model with previously presented ones. The first point that requires careful attention is the comprehensiveness of the model. The proposed model is more comprehensive than all previously presented models because it was developed over a diverse set of 1954 pure compounds belonging to various chemical families.

The second point is related to the root-mean-square error (RMSE) of the presented model, which is 0.285 for the logarithm-based H data. This value is lower than those of the best previously presented models such as the models presented by Lin and Sandler[18] (RMSE of 0.34 over 395 pure compounds), Meylan and Howard[8,9] (RMSEs of 0.52 and 0.42 over the same data set as used by Lin and Sandler[18]), and Modarresi et al.[13] (RMSE of 0.564 over 940 pure compounds).

4. CONCLUSIONS

In this work, a QSPR model was successfully developed to estimate the Henry's constants of pure compounds in water (H) at ambient conditions. The model was developed using 1954 pure compounds belonging to diverse chemical families. These 1954 pure compounds represent many chemical families; therefore, the model has a wide range of applicability, but application of the model is restricted to compounds similar to those used to develop the model. Its application to compounds that are completely different is not recommended. More meticulous investigations should be performed on this concept.

■ ASSOCIATED CONTENT

Supporting Information

Table containing the names, selected molecular descriptors, and predicted values for all compounds used in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*(F.G.) E-mail: fghara@gmail.com. Fax: +98 21 77926580. (A.H.M.) Email: amir-hossein.mohammadi@mines-paristech.fr. Phone: + (33) 1 64 69 49 70. Fax: + (33) 1 64 69 49 68.

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Harrison, R. M. *Pollution: Causes, Effects and Control*; Royal Society of Chemistry Publishing: London, 2001.

(2) Morrison, R. D. *Environmental Forensics: Principles & Applications*; CRC Press: Boca Raton, FL, 1999.

(3) Lyman, W. R. *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*; American Chemical Society: Washington, DC, 1990.

(4) MacKay, D.; Shiu, W. S.; Ma, K. C.; Boethling, R. S. Henry's law constant. In *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*; Boethling, R. S., Mackay, D., Eds.; Lewis: Boca Raton, FL, 2000; pp 69−87.

(5) Staudinger, J.; Roberts, P. V. A critical review of Henry's law constants for environmental applications. *Crit. Rev. Environ. Sci. Technol.* **1996**, *26* (3), 205−297.

(6) Hine, J.; Mookerjee, P. K. The intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. *J. Org. Chem.* **1975**, *40* (3), 292−298.

(7) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Empirical method for estimation of Henry's law constant of

non-electrolyte organic compounds in water. *J. Chem. Thermodyn.* **2012**, *47*, 295−299.
(8) Meylan, W. M.; Howard, P. H. Bond contribution method for estimating Henry's law constants. *Environ. Toxicol. Chem.* **1991**, *10* (10), 1283−1293.
(9) Meylan, W. M.; Howard, P. H. *HENRYWIN*, version 3.10; Syracuse Research: Syracuse, NY, 2000.
(10) Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *J. Chem. Soc., Perkin Trans. 2* **1994**, No. 8, 1777−1791.
(11) Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR study of the solubility of gases and vapors in water. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 1162−1168.
(12) Dearden, J. C.; Ahmad, S. A.; Cronin, M. T. D.; Sharra, J. A.; Gundertofte, K.; Jørgensen, F. S., 273-274.
(13) Dearden, J. C.; Cronin, M. T. D.; Sharra, J. A.; Higgins, C.; Boxall, A. B. A.; Watts, C. D.; Schüürmann, G. F. The Prediction of Henry's Law Constant: A QSPR from Fundamental Considerations. In *Quantitative Structure−Activity Relationships in Environmental Science VII*; Chen, F., Schuurmann, G., Eds.; SETAC Press: Pensacola, FL, 1997; pp 135−142.
(14) English, N. J.; Carroll, D. G. Prediction of Henry's Law Constants by a Quantitative Structure−Property Relationship and Neural Networks. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3−6), 1150−1161.
(15) Yao, X.; Liu, M.; Zhang, X.; Hu, Z.; Fan, B. Radial basis function network-based quantitative structure−property relationship for the prediction of Henry's law constant. *Anal. Chim. Acta* **2002**, *462* (1), 101−117.
(16) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Tiné, M. R. Quantitative Structure−Property Relationship (QSPR) Prediction of Solvation Gibbs Energy of Bifunctional Compounds by Recursive Neural Networks. *J. Chem. Eng. Data* **2010**, *55* (12), 5425−5428.
(17) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tine, M. R. Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. *J. Chem. Inf. Model.* **2006**, *46* (5), 2030−42.
(18) 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.
(19) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A fuzzy ARTMAP-based quantitative structure−property relationship (QSPR) for the Henry's Law constant of organic compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (1), 85−112.
(20) 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* (11), 2067−2076.
(21) 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* (20), 10149−10152.
(22) Yaws, C. L. *Yaws' Handbook of Thermodynamic and Physical Properties of Chemical Compounds*; Knovel: Norwich, NY, 2003.
(23) *HyperChem Release 7.5 for Windows*; Hypercube, Inc.: Gainesville, FL, 2002.
(24) Talete srl, Dragon for Windows (Software for Molecular Descriptor Calculation), version 5.5, 2007.
(25) Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267−281.
(26) Todeschini, R., Consonni, V., Eds. *Molecular Descriptors for Chemoinfomatics: Volume I: Alphabetical Listing, Volume II: Appendices, References*; Methods and Principles in Medicinal Chemistry Series; Wiley: New York, 2009; Vol. *41*.
(27) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: Multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **2004**, *515* (1), 199−208.
(28) Gharagheizi, F.; Sattari, M. Prediction of triple-point temperature of pure components using their chemical structures. *Ind. Eng. Chem. Res.* **2010**, *49* (2), 929−932.
(29) Gharagheizi, F.; Sattari, M. Prediction of the θ(UCST) of polymer solutions: A quantitative structure−property relationship study. *Ind. Eng. Chem. Res.* **2009**, *48* (19), 9054−9060.
(30) Gharagheizi, F. Chemical structure-based model for estimation of the upper flammability limit of pure compounds. *Energy Fuels* **2010**, *24* (7), 3867−3871.
(31) Gharagheizi, F. A QSPR model for estimation of lower flammability limit temperature of pure compounds based on molecular structure. *J. Hazard. Mater.* **2009**, *169* (1−3), 217−220.
(32) Gharagheizi, F. Prediction of upper flammability limit percent of pure compounds from their molecular structures. *J. Hazard. Mater.* **2009**, *167* (1−3), 507−510.
(33) Gharagheizi, F.; Mirkhani, S. A. Predictive Quantitative Structure−Property Relationship Model for the Estimation of Ionic Liquid Viscosity. *Ind. Eng. Chem. Res.* **2012**, *51* (5), 2470−2477.
(34) Gharagheizi, F.; Eslamimanesh, A.; Sattari, M.; Mohammadi, A. H.; Richon, D. Corresponding States Method for Determination of the Viscosity of Gases at Atmospheric Pressure. *Ind. Eng. Chem. Res.* **2012**, *51* (7), 3179−3185.
(35) Gharagheizi, F.; Eslamimanesh, A.; Sattari, M.; Mohammadi, A. H.; Richon, D. Corresponding States Method for Evaluation of the Solubility Parameter of Chemical Compounds. *Ind. Eng. Chem. Res.* **2012**, *51* (9), 3826−3831.
(36) Gharagheizi, F. Determination of Diffusion Coefficient of Organic Compounds in Water Using a Simple Molecular-Based Method. *Ind. Eng. Chem. Res.* **2012**, *51* (6), 2797−2803.
(37) Gharagheizi, F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40* (1), 159−167.

NOTE ADDED AFTER ASAP PUBLICATION

After this paper was published online March 15, 2012, a correction was made to the unit of *H* in the Introduction section. The corrected version was reposted March 19, 2012.