QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors

Cikui Liang[§]

Department of Environmental Science & Engineering, Oregon Graduate Institute, P.O. Box 91000, Portland, Oregon 97291-1000

David A. Gallagher*

Oxford Molecular Group, 14940 NW Greenbrier Parkway, Beaverton, Oregon 97006-5733

Received November 3, 1997

To date, most reported quantitative structure−property relationship (QSPR) methods to predict vapor pressure rely on, at least, some empirical data, such as boiling points, critical pressures, and critical temperatures. This limits their usefulness to available chemicals and incurs the time and expense of experimentation. A model to predict vapor pressure from only computationally derived molecular descriptors, allowing study of hypothetical structures, is described here. Several multilinear regressions and artificial neural network analyses were tested with a range of descriptors (e.g., topological and quantum mechanical) derived solely from computations on molecular structure data. From a set of 479 compounds, a linear regression with an r^2 of 0.960 was achieved using polarizibility and polar functional group counts as descriptors. This new computationally based model also proves to be more accurate and works over a wider range of compound classes than most previously reported models.

INTRODUCTION

Vapor pressure plays an important role in the transport, distribution, and fate of environmental pollutants in the atmosphere and hence, the environmental acceptability of chemical products and processes.[1] Experimental vapor pressure measurement of the ever-growing number of actual and potential chemical products is both time-consuming and expensive. On the other hand, a statistically significant quantitative structure−property relationship (QSPR) that requires only chemical structure data could serve as a tool to predict reliable vapor pressure data in a fraction of the time and expense. To date, most reported QSPR methods to predict vapor pressure still depend on some empirical data, such as boiling points, critical pressures, and critical temperatures.[2−5] However, a QSPR approach that requires no empirical data (other than experimental vapor pressure values to develop the model) could facilitate extensive prescreening of hypothetical chemical products, and so obviate the need for prior synthesis and testing.

When a structure−property relationship is found, it might also provide insight into which aspect of the molecular structure influences the property. Such insight may facilitate a systematic approach to design of new molecules with more desirable properties. The development of a QSPR using only computationally derived descriptors has already been successfully applied to the prediction of water solubility.[6] This approach has the potential to be applied to many other chemical and physical properties.

MODEL DEVELOPMENT

The QSPR models were developed with the CAChe Worksystem 3.8 (Oxford Molecular Group Inc., Beaverton, Oregon) running on a Power Macintosh 8500 and Tsar 3.0 (Oxford Molecular Group Inc., Beaverton, Oregon) running on a Silicon Graphics Indigo R3000-based system. The CAChe Worksystem performs calculations of various physical, chemical, topological, and electronic descriptors directly from the structure of a compound. Tsar provides statistical analysis of the numerical data, including linear regression and artificial neural network analysis, to develop and optimize the QSPRs.

All subcooled vapor pressure (log p_L) data used in the model development were obtained directly from the literature.[7] Where different values for the same compound were found, the average of the reported values was used. Vapor pressures reported at temperatures other than 25 °C were extrapolated to 25 °C using the approach detailed in previous reports.[8,9] Experimental log p_L data for 479 compounds were collated. Compound classes included acids, alcohols, aldehydes, alkanes, alkenes, alkynes, amines, aromatic compounds and polynuclear aromatic hydrocarbons (PAHs), dibenzofurans, dioxins, ethers, esters, ketones, nitriles and other nitrogen-containing compounds, polychlorinated biphenyls (PCBs), and sulfur-containing compounds. Full information on all compounds has not been included here due to the size of the dataset. However, it is available in electronic form from the authors.

Structures of all the molecules were sketched in a ProjectLeader table using the CAChe Editor, and their geometries were then optimized using the default "standard procedures" based on CAChe MOPAC PM3.[10] Various properties, parameters, and molecular descriptors were selected directly from a menu of over 100 functions provided in the CAChe Worksystem. Selections were based on the results from previous work in the literature and any properties that might be related to vapor pressure. A list of the 25

* Corresponding author.
§ Current address: Oxford Molecular Group, Inc., Beaverton, OR.

S0095-2338(97)00289-8 CCC: $15.00
Published on Web 02/19/1998

Table 1. Descriptor List

class	molecular descriptor
constitutional descriptors	molecular weight, halogenated atom counts, and functional group counts (amine group, carbonyl group, carboxylic acid group, hydroxyl group, nitrile group, nitro group)
electrostatic descriptors	partial charges, charged surface areas
topological descriptors	Kier and Hall connectivity indices (zero order, first order, and second order) and valence connectivity indices (zero order, first order, and second order)
geometrical descriptors	solvent accessible surface, isosurface volume, molar volume
quantum-chemical descriptors	dipole moment, quadrupole moment, polarizibility, HOMO and LUMO energies, dielectric energy

descriptors considered in the model development, based on the classification of Katritzky et al.,[11] is given in Table 1.

For each compound, values of the 25 descriptors were calculated using the CAChe WorkSystem and then transferred to Tsar along with the experimental data for detailed statistical analysis. Principal component analysis (PCA) was used to identify highly correlated descriptors, so that regressions could be reduced to a minimum number of descriptors without losing significant information.[12] Linear regression (simple-, multi-, and stepwise-) and neural network analyses[13] (Forward Feed) were then carried out to derive correlations between log p_L and the descriptors that survived PCA. Both the regression and the neural net analyses employed the default settings in Tsar except where specified otherwise. The regression analyses were carried out on individual compound classes as well as on the complete data set of 479 compounds. Cross validation of the regression is performed automatically by Tsar in the following manner.[12] Two-thirds of the data is used to develop a regression to predict the vapor pressure of the remaining third that is not used in the regression. This procedure is repeated for each third and the reported cross validation coefficient (r_{cv}^2) is the average of all three. Cross validation is used to give an estimate of the true predictive power of the model, i.e., how reliable the predicted values for untested compounds are likely to be. A regression coefficient corrected for the degrees of freedom (r_f^2) was also calculated according to[14]

$$r_f^2 = r^2 - \frac{p - (1 - r^2)}{n - p - 1} \quad (1)$$

where r^2 is the regression coefficient, p is the number of descriptors in the QSPR model, and n is the total number of molecules. Finally, r_{cv}^2, r^2, r_f^2, standard errors (s), and average unsigned errors (E_A) from different regression equations were compared. The selected QSPRs were based on the highest r_{cv}^2 and lowest s.

It has been reported that artificial neural networks (NN) sometimes provide more accurate estimates than multilinear regression.[15] Hence, NN analyses were also carried out with the same descriptor sets that were used in the regression analyses. In each NN analysis, 30% of the 479 compounds were randomly set aside as a test set to examine the predictive power of the network during development. All the NN analyses were carried out with one hidden layer and one output node (log p_L). Depending upon the number of input nodes and data points, different numbers of hidden nodes were used in order to obtain acceptable residual variance values.[16]

RESULTS AND DISCUSSION

The vapor pressure of a compound is related to the forces of intermolecular attraction.[17] The stronger the intermolecular forces are, the more tightly the molecules are held together in a condensed phase and, hence, the lower the vapor pressure will be. Molecular interactions in most nonionic liquids can be divided into three types: dipole−dipole, dipole−induced dipole, and induced dipole−induced dipole or dispersion forces.[17] Hydrogen bonding may also be present. Clearly, dipole interactions are related to the dipole moment of a whole molecule or of a part of a molecule, such as a functional group, e.g., carbonyl. The dispersion forces are a function of the molecule's polarizability (α),[18] while hydrogen bonding can be facilitated by the presence of −OH, −NH, or −SH groups, etc.

The correlation of each descriptor alone was tested against the log p_L. Polarizability provided the best correlation with log p_L for the complete set of 479 compounds. As shown in Table 2, polarizability also produced relatively high r^2 values within many of the individual compound classes and did especially well for the nonpolar hydrocarbons with $r^2 = 0.997$. This is consistent with the fact that the polarizability is related to the dispersion forces or induced dipole−induced dipole interactions, which are the main component of the intermolecular forces in nonpolar compounds.[17] On the other hand, polarizability showed lower correlations for relatively polar compounds, such as the alcohols, amines, and halogenated ketones. This may be due to the potential for hydrogen bonding and/or dipole (electrostatic) interactions which are not adequately accounted for by the polarizability descriptor.

To improve the regression model for all classes of compounds, additional descriptors were tested to accommodate the additional interactions present in polar and in hydrogen-bonding compounds. Hence, various other descriptors were tested in conjunction with the polarizability. Polar functional group counts that included −OH, −NH, $-NO_2$, −C≡N, −COOH, and >C=O gave the best improvement in r^2. The regression correlation employing polarizability and the six functional group counts is given by

$$\log p_L = -0.432\alpha - 1.382(\mathrm{OH}) - 0.482(\mathrm{C{=}O}) - 0.416(\mathrm{NH}) - 2.197(\mathrm{COOH}) - 1.383(\mathrm{NO_2}) - 1.101(\mathrm{C{\equiv}N}) + 4.610 \quad (2)$$

$$r^2 = 0.960, \quad r_{cv}^2 = 0.957, \quad s = 0.534, \quad n = 479$$

The addition of the six polar functional group counts improves the r^2 from 0.922 to 0.960 for the set of 479 compounds (r_f^2 from 0.920 to 0.945), and the r_{cv}^2 rises from 0.920 for the single-descriptor to 0.957 for the seven-descriptor model. Presumably, the functional groups account for many of the dipole−dipole, dipole−induced dipole, and

Table 2. Regression Results with Single Descriptor, Polarizibility (α)

compd class	regression eq	n^a	r^2	r_{cv}^2	s
nonpolar alkanes, alkenes, and alkynes	$\log pL = -0.541\alpha + 5.600$	70	0.997	0.997	0.185
benzene and its nonpolar derivatives	$\log pL = -0.485\alpha + 5.428$	13	0.992	0.990	0.069
aldehydes	$\log pL = -0.443\alpha + 4.127$	7	0.990	0.986	0.107
dioxins and furans	$\log pL = -0.500\alpha + 5.615$	24	0.968	0.964	0.393
PAHs	$\log pL = -0.414\alpha + 4.530$	20	0.959	0.950	0.365
acids	$\log pL = -0.582\alpha + 2.991$	9	0.957	0.920	0.207
nitriles	$\log pL = -0.384\alpha + 3.251$	11	0.956	0.933	0.194
esters	$\log pL = -0.422\alpha + 4.289$	39	0.949	0.944	0.202
ethers	$\log pL = -0.393\alpha + 4.128$	24	0.922	0.913	0.478
pcbs	$\log pL = -0.352\alpha + 3.456$	27	0.877	0.823	0.528
ketones	$\log pL = -0.419\alpha + 3.947$	18	0.876	0.840	0.258
alcohols	$\log pL = -0.353\alpha + 2.643$	58	0.782	0.776	0.398
amines	$\log pL = -0.417\alpha + 4.107$	30	0.760	0.731	0.580
halogenated alkanes	$\log pL = -0.472\alpha + 5.341$	49	0.731	0.701	0.517
general model for all compds in the database	$\log pL = -0.401\alpha + 3.940$	479	0.922	0.920	0.745

[a] Number of data points.

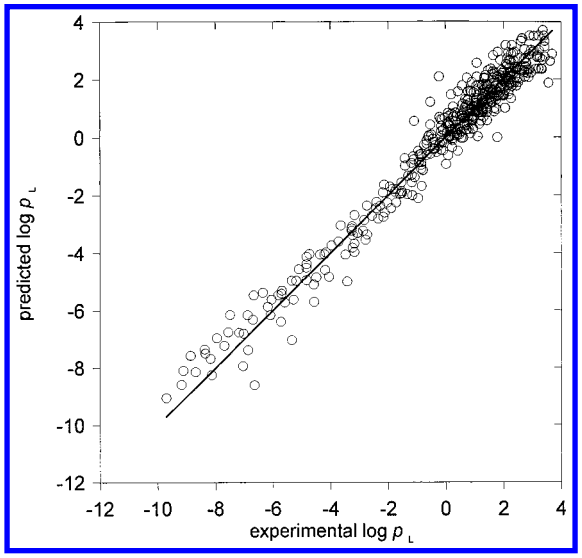

Figure 1. Predicted log p_L vs experimental log p_L.

hydrogen bonding interactions of the polar compounds considered here. A plot of the predicted log p_L calculated with eq 2 vs the experimental log p_L is shown in Figure 1.

COMPARISON OF RESULTS

To further test the predictive power of the model, some compounds that gave large errors with another reported QSPR model [4] were evaluated, and the results are given in Table 3. The predicted log p_L values using this model (eq 2) show higher accuracy for the cited compounds than those obtained by Banerjee et al.[4] A recent paper by Basak et al.[23] described the prediction of vapor pressure from only computationally derived descriptors, but without the use of any quantum mechanics-based calculations, such as polarizability. However, a lower r^2 value was achieved from their study as compared to the r^2 value obtained in this study, i.e., 0.84 versus 0.96. Equation 2 was also used to estimate log p_L values for some additional compounds that were not included in the original regression model, and the predicted log p_L values are also reported in Table 3. The accuracy of the predicted log p_L values for both polar and nonpolar compounds with this QSPR model is generally better than previously reported QSPR models.[4,23]

Table 3. Estimated log p_L for Selected Organic Compounds

compds	Exptl log p_L (Torr)	predicted log p_L (eq 2 in this work) (Torr)	error (eq 2 in this work)	reported error[h]
diisopropyl ether[a,b]	2.176	1.578	−0.598	−1.26
1,1,1-trichloroethane[a,b]	2.093	2.039	0.054	−1.24
hexane[a,b]	2.180	1.855	−0.325	−1.11
formic acid[a,b]	1.635	1.538	−0.097	1.08
propanoic acid[a,b]	0.618	0.641	0.023	−1.29
ethylene glycol[b,c]	−1.036	−0.605	−0.431	−2.19
quinoline[b,d]	−1.018	−0.907	−0.111	
isoquinoline[b,e]	−1.197	−0.855	0.342	
nicotine[b,f]	−1.372	−1.416	−0.044	
quinaldine[b,c]	−1.379	−1.248	0.140	
N-ethylcarbazole[b,d]	−3.224	−3.639	−0.415	
dimethyl disulfide[b,f]	−1.457	−1.469	−0.012	
methyl *tert*-butyl ether[b,g]	2.389	2.080	−0.309	

[a] Compounds included in the training set for model development. [b] From ref 4. [c] Compounds not included in the training set for model development. [d] From ref 19. [e] From ref 20. [f] From ref 21. [g] From ref 22. [h] Obtained from Table 2 in Ref 4.

Table 4. Summary of Results for Multiple Linear Regression (MLR) an Neural Network (NN) Analyses

method	no. of descriptor	NN confgrtn	r^2	r_f^2	r_{cv}^2	s	av unsigned error
MLR	7		0.960	0.945	0.957	0.534	0.396
NN	7	7-5-1[a]	0.961		0.960	0.522	0.386
NN	25	25-9-1	0.973		0.973	0.437	0.325

[a] 7, 5, and 1 are the number of input nodes, hidden nodes, and the output nodes, respectively.

NEURAL NETWORKS

NN analyses are claimed to be superior to linear regression in their ability to handle nonlinear correlations.[15] Hence, NN analyses[13] were tested on several of the vapor pressure regression models. The regression coefficients of the NN predicted values versus experimental were then compared to those from linear regression in Table 4. Since 30% of the compounds were selected randomly as a test set at the beginning of each neural net analysis, the values for r^2 and r_{cv}^2 vary with each run depending on the compounds included in the training set. Each analysis was repeated several times to find the best values of r^2 and r_{cv}^2.

The results in Table 4 indicate that NN analysis showed no significant advantage over linear regression for this vapor pressure prediction model. Optimization of the neural networks may have provided better results; however, this was beyond the scope of this work.

ERRORS

Undoubtedly, a component of the residual error in the final QSPR models is attributable to variance in the experimental data used to develop the regression. However, the component of experimental uncertainty could not be estimated as most of the reported vapor pressure data did not include experimental errors. It is likely that some of the chlorinated PCB outliers may be subject to larger than average experimental errors because of the difficulty of measuring their extremely low vapor pressures.

In both the linear regression and neural net QSPR models, at least 80% of the outliers having greater than 1 log unit error are compounds containing sulfur or halogens. However, sulfur- and halogen-containing compounds only account for 30% of the total test set. These elements are reported to have less reliable parameterization in MOPAC PM3 than carbon, hydrogen, oxygen, and nitrogen.[10] Hence, the larger errors in predicted vapor pressure for sulfur and halogen containing compounds are consistent with the fact that the key polarizability descriptor was calculated by MOPAC PM3. Including separate atom counts of S, I, Cl, F, etc. as additional descriptors failed to improve the results. The authors plan to investigate using higher accuracy *ab initio* calculations in place of the semiempirical MOPAC program. *ab Initio* may provide more accurate polarizability calculations for halogenated and sulfur compounds which could improve the r^2 further.

CONCLUSION

The multilinear regression model based on a simple semiempirical calculation and functional group counts provides reasonably accurate vapor pressure predictions ($r^2 = 0.960$) across a wide range of compound classes. The average predicted log p_L values are more accurate than some previously reported models.[4,23]

Unlike previously reported methods, the QSPR models described here have the advantage of not requiring any experimental parameters and, hence, can be applied to hypothetical compounds across a broad range of compound classes.

Both the polarizability and polar-functional-group-count descriptors used in this model can be rationalized in terms of the intermolecular forces that influence vapor pressure.

A significant part of the standard error is presumably due to the unreported errors in the experimental data used to calibrate the QSPR. For some compounds, the poor parametrization of sulfur and the halogens in MOPAC PM3 may also contribute to the errors.

ACKNOWLEDGMENT

The authors wish to express their appreciation to Drs. Vijay Gombar, George Purvis, Nigel Richards, and Timothy Clark for their advice and guidance in the preparation of this manuscript.

Supporting Information Available: Table of comparison of experimental log p_L and predicted log p_L using seven descriptor multilinear regression model (12 pages). See any current masthead page for ordering and Web access instructions.

REFERENCES AND NOTES

(1) Shiu, W. Y.; Doucette, W.; Gobas, F. A. P. C.; Andren, A.; Mackay, D. *Environ. Sci. Technol.* **1988**, *22*, 651−658.
(2) Mackay, D. M.; Bobra, A.; Chan, D. W.; Shiu, W. Y. *Environ. Sci. Technol.* **1982**, *16*, 645−649.
(3) McGarry, *J. Ind. Eng. Chem. Process Des. Dev.* **1983**, *22*, 313−322.
(4) Banerjee, S.; Howard, P. H.; Lande, S. S. *Chemosphere* **1990**, *21*, 1173−1180.
(5) Mishra, D. S.; Yalkowsky, S. H. *Ind. Eng. Chem. Rev.* **1991**, *30*, 1609−1612.
(6) Liang, C.; Gallagher, D. A. *Am. Lab*. **1997**, March, 34−40.
(7) Mackay, D.; Shiu, W. Y.; Ma, K. C. *Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals;* Lewish Publisher: Boca Raton, 1992.
(8) Ohe, S. *Computer Aided Data Book on Vapor pressure;* Data Book Publishing Company: Tokyo, 1976.
(9) Lide, D. R. *CRC Handbook of Physics and Chemistry*, 75th ed.; CRC Press: Boca Raton, 1994.
(10) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 221−264.
(11) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, 279−287.
(12) *Tsar User Guide Issue 6.0*; Oxford Molecular Ltd.: Oxford, England, 1995.
(13) Good, A. C.; So, S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrixes. *J. Med. Chem.* **1993**, *36*, 433−438
(14) *BMDP Statistical Software Manual*; Dixon, W. J., Ed.; University of California Press: Los Angeles, 1992; Vol. 1, pp 387−425.
(15) Livingstone, D. J.; Salt, D. W. *Bioorg. Med. Chem. Lett*. **1992**, *2*, 213-218.
(16) Andrea, R. A.; Kalayeh, H. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(17) Schwarzenbach, R. P.; Gschwead, P. M.; Imboden, D. M. *Environmental Organic Chemistry*; John Wiley & Son, Inc.: New York, 1993.
(18) Atkins, R. W. *Physical Chemistry*; W. H. Freeman Company: San Francisco, 1978.
(19) Van de Rostyne, C.; Prausnltz, J. M. *J. Chem. Eng. Data* **1980**, *25*, 1−3.
(20) Das, a.; Frenkel, M.; Gadalla, N. A. M.; Kudchadker, S.; Marsh, K. N.; Rodger, A. S.; Wilhoit, R. C. *J. Phys. Chem. Ref. Data* **1993**, *22*, 659−782.
(21) Timmermans, J. *Physico-Chemical Constants of Pure Organic Compounds*; Elsevier Publishing Co.: New York, 1950.
(22) Budavari, S.; O'Neil, N. J.; Smith, A.; Heckelman, P. E. *The Merck Index*, 11th ed.; Merck & Co., Inc.: Rahway, NJ, 1989.
(23) Basak, S. C.; Gute, B. D.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 651−655.

CI970289C