**Quantitative structure-property relationship studies for prediction vapor pressure of volatile organic compounds**

MOUNIA ZINE1, AMEL BOUAKKADIA1,2, LEILA LOURICI3, DJELLOULMESSADI1[[1]](#footnote-1)\*

*1Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP. 12, 23000 Annaba, Algeria*

*2Abbes Laghrour University, Faculty of sciences and technology, Khenchela, Algeria.*

*3Chadeli Ben djedid University, BP.73, 3600 El taref, Algeria*

*Abstract:* A theoretical model (QSPR) using multiple linear regression analysis for predicting vapor pressure (*p*v / Pa) of volatile organic compounds (VOCs) has been developed, a series of 51compounds were analyzed by multiple linear regression analysis. At first, the data set was separated arbitrarily into a training set (39 chemicals) and a test set (12 chemicals) for statistical external validation. The four-dimensional Model was developed using as independent variables theoretical descriptors derived from DRAGON software when applying GA (Genetic Algorithm)-VSS (Variable Subset Selection) procedure. The obtained model was used to predict the vapor pressure of test set compounds, and an agreement between experimental and predicted values was verified. This model, with high statistical significance (*R*2 = 0.9090, *Q*²Loo = 0.8748, *Q*²Ext = 0.8307,*s* = 0.24), could be used adequately for the prediction and description of the log (*p*v / Pa) of other VOCs. The applicability domain of MLR models was investigated using William‘s plot to detect outliers and outsides compounds.

*Keywords:* Molecular descriptors; VOCs; log (*p*v */* Pa); Multiple Linear Regression.

QSPR STADIES OF VAPOR PRESSURE

INTRODUCTION

Volatile organic compounds (VOCs) are molecules which can contain H and C atoms but also other elements such as O, N, Cl, F, P, S,... and metals and/or metalloids, and which are almost entirely state of vapor under normal conditions of temperature and pressure. They include 210 species and 23 large families. These compounds can be of natural origin (terpenes) but very often they are contaminants mainly of human activity.1

The sectors of activities the more strongly transmitters of VOCs are road transport, industry, agriculture, and the tertiary sector. The other air pollutants that can be cited are the biological contaminants (bacteria, pollen, fungi), the physical contaminants (metals, particles, dust, the radioactivity), and chemical contaminants which are part of the VOCs with gases (CO, O3, NOx, SO2, fluorocarbons), dioxins and furans.1

Experience is a direct way to obtain the activity data for organic compounds, which have many shortcomings, such as the need for large test organisms, high costs, long time duration, and value difference measured between different researchers. Consequently, it would be impossible to test the activity values of all organic compounds by experiments.

As new compounds are springing up, other difficulties will also arise. Therefore, it is necessary to use the theoretical methods to waiver the disadvantages of the experiment and to predict the data of compounds exactly.

With the rapid development of computer science and theoretical quantum chemical studies, one can speedily and precisely obtain the quantum chemical parameters of compounds by computation. These structural parameters along with the introduction of the quantitative structure-activity relationship (QSAR) models can increase the interpretability and predict the activity of new organic compounds.2

Quantitative structure-property relationships (QSPR) have gained wide attention in the area of separation science recently. These models are based on the relationship between structures and property of compounds.3

Vapor pressure of different compounds can be predicted from their formula and even unknown compounds can be identified by using this method. In general, QSPR models attempt to predict the vapor pressure of a molecule by characterizing it with a series of molecular descriptors. These models can effectively be used for the prediction of molecular structures, determination of vapor pressure.

The aim of this study is to found a statistical model for the prediction of vapor pressures of some volatiles organic compounds. We have validated the model by dividing the data set arbitrarily into training (39 compounds) and test set (12 compounds).

Different statistical techniques were used to develop the model to highlight the structural requirements for ideal vapor pressure. The three objectives of the present paper have bee First, to explore the structure-activity relationships of vapor pressure of diverse volatiles organic compounds. Second, to select the best predictive model from among all comparable chemometric models for the property and Third, verification of the performance and stability of the obtained model by two approaches (MLR). The model obtained shows which descriptors play a significant role in log (*p*v / Pa) variation of these compounds.

EXPERIMENTAL

*Data set*

The (*p*v / Pa) experimental values of 51 selected, structurally heterogeneous VOCs were collected from previous works, 4 converted to log (*p*v / Pa) values.

*Molecular descriptors generation*

The structures of the molecules were drawn using Hyperchem 6.03 software.5 The final geometries were obtained with the semi empirical method PM3. All calculations were carried out at the RHF (restricted Hartree–Fock) level with non configuration interaction. The molecular structures were optimized using the algorithm Polak-Ribiere and a gradient norm limit of 0.01 kcal. Ǻ-1. mol-1. The resulted geometry was transferred into the software Dragon version 5.5 to calculate 1600 descriptors of the type Geometrical and GETAWAY (Geometry, Topology and Atoms Weighted Assembl Y) 6.

Descriptors with constant or near constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient *r* ≥ 0.95), the one showing the highest pair correlation with the other descriptors was excluded. The GA (Genetic Algorithm)7 has been considered superior to other methods of variable selection techniques. So, variable selection was performed on the training set, using GA in the MobyDigs version of Todeschini8 by maximizing the cross-validated explained variance *Q*²LOO.

*Model Development and Validation*

Models with four variable were performed by the software MOBYDYGS.8 The goodness of fit of calculated models were assessed by means of the multiple determination coefficients, *R*2 and the standard deviation error in calculation (SDEC).



Cross validation techniques allow the assessment of internal predictivity (*Q*2LMO cross validation; bootstrap) in addition to the robustness of model (*Q*2LOO cross validation). Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out. This procedure is repeated for all compounds of the training set, obtaining a prediction for everyone. If each compound is taken away one at a time the cross validation procedure is called leave-one-out technique (LOO), otherwise leave-more-out technique (LMO).

(LOO) or (LMO)correlation coefficient, generally indicated with *Q*2, is computed by evaluating the accuracy of these “test” compounds prediction.



The “hat” of the variable *y*, as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index “*i*/*i*” indicates that the predicted values come from models built without the predicted compound. TSS is the total sum of squares. The predictive residual sum of squares (PRESS) measures the dispersion of the predicted values. It is used to define *Q*2and the standard deviation error in prediction (SDEP).



A value *Q*2> 0.5 is generally regarded as a good result and *Q*2> 0.9 as excellent.9,10 However, studies 11,12 have indicated that while *Q*2 is a necessary condition for high predictive power a model, is not sufficient. To avoid overestimating the predictive power of the model *LMO* procedure (repeated 5000 times, with 4 objects left out at each step) was also performed (*Q2* L(4)O).

In bootstrap validation technique *K* n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then *Q*2 is calculated for each model. The bootstrapping was repeated 8000 times for each validated model. By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of *Q*2Ext, which is defined as :



Here *n*Ext and *n*tr are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively. The data set was divided arbitrary into a training set (39 objects) used to develop the QSAR models and a validation set (12 objects), used only for statistical external validation. Other useful parameters are *R*2, calculated for the validation chemicals by applying the model developed on the training set, and external standard deviation error of prediction (SDEPExt), and defined as:



Where the sum runs over the test set objects (next). According to Golbraikh and Tropsha.12 A QSPR model is successful if it satisfies several criteria as follows:









Here











where is the correlation coefficient between the calculated and experimental values in the test set; *r*²0 (calculated versus observed values) and *r*’²0 (observed versus calculated values) are the coefficients of determination; *k* and *k’* are slopes of regression lines through the origin of calculated versus observed and observed versus calculated, respectively *yr*0i,*r*0i; are defined as *yr*0i=*k*i and * r*0i=*k*’*y*i and the summations runs over the test set.

The AD (Applicability Domain) was discussed by the Williams plot 8, 9 of jacknified residuals versus leverages (hat diagonal values (*h*i). The jacknifed residuals (or Studentized residuals) are the standardized cross-validated residuals. Each residual is divided by its standard deviation, which is calculated without the *i*th observation. The leverage (*h*i) value of a chemical in the original variable space is defined as:



Where xi is the descriptor row-vector of the query compound and X is the *n*(*p*+1) matrix of p model parameter values for n training set compounds. The superscript T refers to the transpose of the matrix/vector. The warning leverage value (*h*\*) is defined as 3(*m*+1)/*n*. When *h* value of a compound is lower than *h*\*, the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with *h*i>*h*\*will reinforce the model if the chemical is in the training set. But such a chemical in the validation set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residual may be low. Thus the leverage and the jacknified residual should be combined for the characterization of the AD.

In this stage, linear QSPR model was developed and evaluated to predict the log (*p*v / Pa) of the compounds. The study we conducted consists of the multiple linear regressions (MLR) available in the MobyDygssoftware.

RESULTS AND DISCUSSION

In order to predict log (*p*v / Pa) application of the GA-VSS lead to several good models for the prediction based on different sets of molecular descriptors.

The best model obtained using 39 compounds is a four-dimensional model (*X*0*sol*, *SpPosA\_H*2, *GATS*2*e* and *Hy*) with a high predictive power.

The Multiple Linear Regression model (MLR) is given by:

Log *(p*v */* Pa*)* = 11.00 – 0.46 *X*0*sol* – 12. 30 *SpPosA\_H*2 + 1.14 *GATS*2*e* – 1.23 *Hy*  (16)

Statistical parameters for the model with: ntr= 39 nExt= 12 are:

*R*² = 90.09% *Q*²Loo = 87.48% *Q*²Boot= 85. 55% *F* = 77. 25 *Q*2Ext= 83. 07%

SDEP= 0.256 SDEC= 0. 227 SDEPExt = 0. 297 *s* = 0.24 *K*xx= 38.51 *K*xy *=* 45.57

The reported fitting and validation parameters have, as expected, high values indicating that the model has very good predictive performance and the descriptors involved in it well describe the vapor pressure.

The high absolute *t*-values shown in TABLE I. express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The *t*-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (descriptors interactions).

Descriptors with t-probability values below 0.05 (95 % confidence)) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance.13 The smaller t- probability suggests the more significant descriptor. The t-probability values of four descriptors are very small, indicating that all of them are highly significant descriptors.

The VIF Table I. is uniform and equal to 1. 00 if there is no linear correlation between a given variable and rest of the variables in the regression equations. Higher values of VIF indicate a more serious multi-co linearity problem.

Models would not be accepted if they contain descriptors with VIFs above a value of five.14



Where *R*²j is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model.15

TABLE I. Characteristics of the selected descriptors in the best GA/MLR model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Descriptors | Coefficient | Standard error coefficient | *t* | t-Probability | VIF |
| Constant | 11.0490 | 0.6288 | 17.57 | 0.000 |  |
| *X*0*sol* | -0.4602 | 0.0460 | -9.81 | 0.000 | 1.720 |
| *SpPosA\_H*2 | -12.3320 | 1.2100 | -10.19 | 0.000 | 1.528 |
| *GATS*2*e* | 1.1372 | 0.1423 | 7.99 | 0.000 | 1.074 |
| *Hy* | -1.2333 | 0.1275 | -9.68 | 0.000 | 1.944 |

*Applicability Domain*

On analyzing the model applicability domain from Williams plot, all residuals were located within the range of three standard deviations, and there is no influential compound both for training or prediction set Fig. 1, which means that the model has a good external predictivity.

C:\Users\dell\Downloads\Mounia\Figures\Williams plot jackknifed residuals and leverages.tif

**Fig. 1.** Williams plot: jackknifed residuals and leverages.

Fig. 2 which represents the diagram of the statistical coefficients *Q*2and *R*2 akes to compare the results obtained for the randomized models (circle) with the starting model (square). It is clear that the statistics obtained for the modified vectors of the log (*p*v / Pa) are smaller than those of the real models; *Q*2 are lower than (<10 %), and for the major part one obtains even *R*2<30 %.

This allows ensuring that the established model has a real base, and is not due arbitrarily.

C:\Users\dell\Downloads\Mounia\Figures\Randomization test associated to previous QSPR model.tif

**Fig. 2.** Randomization test associated to previous QSPR model.

This technique Plot of cross-validation log (*p*v / Pa) versus experimental log (*p*v / Pa) values are shown in Fig. 3 ensures the robustness of a QSPR model.16

**C:\Users\dell\Downloads\Mounia\Figures\Plot of residual vs. experimental log (Vppa) for the entire data set.tif**

**Fig. 3.** Cross-validation *versus* experimental log (*p*v / Pa) for the entire data set.

*Descriptor Contributions Analysis*

Based on a previously described procedure17, 18 the relative contributions of the four descriptors of the model have been determined. It should be noted that the difference in contribution between two descriptors used in the model indicates that all the descriptors are essential to generate the predictive model (Fig. 4).

Descriptors decrease according to the following order:

*SpPosA\_H*2 (27. 7811 %) > *X*0*sol* (26. 6809 %) > *Hy* (25. 7967 %) > *GATS*2*e* (19. 7413 %).

**C:\Users\dell\Downloads\Mounia\Figures\Relative contributions of the selected descriptors to the MLR model.tif**

**Fig. 4.** Relative contributions of the selected descriptors to the MLR model.

TABLE II. Golbraikh and Tropsha12 criteria.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | (*R*2-*R*20 ) / *R*2 < 0.1 | (*R*2-*R*’20 ) / *R*2 < 0.1 | 0.85 < *k* < 1.15 | 0.85< *k’*< 1.15 |
| Value | -0.085 | -0.087 | 0.973 | 1.022 |

CONCLUSION

Multiple linear regressions were used to construct a quantitative structure property relation model of 51 VOCs for their vapor pressures binding propriety. The model was developed by a genetic algorithm selection of theoretical molecular descriptors from among a wide set obtained with several softwares. The data set separated randomly into two subsets of 39 elements for training and 12 for external validation. The MLRmodelhas good stabilities, robustness and predictivity. The chemical applicability domain of the studied MLR model and the reliability of the predictions were verified by the leverage approach. The equation obtained can be used successfully to estimate the vapor pressures for new compounds or for other compounds whose experimental values are unknown.

REFERENCES

1. V. Cirimele, M. Etter, M. Villian, P. Kintez, *Ann. Toxicol. Anal.* **20** (2008) 67 **(**<https://doi.org/10.1051/ata/2009002>)

2. S. [Chtita](https://www.sciencedirect.com/science/article/pii/S1658365515000813" \l "!), R. [Hmamouchi](https://www.sciencedirect.com/science/article/pii/S1658365515000813" \l "!), M. [Larif](https://www.sciencedirect.com/science/article/pii/S1658365515000813" \l "!), M. [Ghamali](https://www.sciencedirect.com/science/article/pii/S1658365515000813#!), M. [Bouachrine](https://www.sciencedirect.com/science/article/pii/S1658365515000813#!), T. [Lakhlifi](https://www.sciencedirect.com/science/article/pii/S1658365515000813#!), *J. Taibah. Univ. Sci.* **10 (**2016) 868 (<https://doi.org/10.1016/j.jtusci.2015.04.007>)

3. J. [Akbar](https://www.ncbi.nlm.nih.gov/pubmed/?term=Akbar%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23203132) , S. [Iqbal](https://www.ncbi.nlm.nih.gov/pubmed/?term=Iqbal%20S%5BAuthor%5D&cauthor=true&cauthor_uid=23203132) , F. [Batool](https://www.ncbi.nlm.nih.gov/pubmed/?term=Batool%20F%5BAuthor%5D&cauthor=true&cauthor_uid=23203132) , A. [Karim](https://www.ncbi.nlm.nih.gov/pubmed/?term=Karim%20A%5BAuthor%5D&cauthor=true&cauthor_uid=23203132) , K.W. Chan, *Int. J. Mol. Sci.* **13** (2012) 15387 (<https://doi.org/10.3390/ijms131115387>)

4. D. Mackay, W. Y. Shiu, K. C. Ma, *S*. C. Lee, *Handbook of physical-chemical properties and environmental fate for organic chemicals Second edition*, CRC Press Inc, Boca Raton, USA, 2006 ([https://doi.org/10.1201/9781420044393](https://doi.org/10.1201/9781420044393" \t "_blank))

5. HyperChem 6.03 Package. Hypercube, Inc., Gainesville, Florida, USA, 1999; software available at : (<http://www.hyper.com>)

6. Talete Srl. Dragon for windows (Software for Molecular Descriptor Calculation) Version 5.5 Milano, Italy, 2007; software available at : (<http://www.talete.mi.it/>)

7. R. Leardi, R. Boggia, M. Terrile, *J. Chemom.* **6** (1992) 267 (<https://doi.org/10.1002/cem.1180060506>)

8. R. Todeschini, D. Ballabio, V. Consonni, A. Mauri, M. Pavan*, MOBYDIGS, Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for windows*, Milano, Italy, 2009 (<http://www.talete.mi.it/>)

9. L. Eriksson, J. Jaworska, A. Worth, M. Mc. Cronin, R.M. Dowell, P. Gramatica, *Environ. Health. Perspect.* **111** (2003) 1361 (<https://doi.org/10.1289/ehp.5758>)

10. A. Tropscha, P. Gramatica, V.K. Gombar, QSAR. *Comb. Sci.* **22** (2003) 70 (<https://doi.org/10.1002/qsar.200390007>)

11. H. Kubinyi, F.A. Hamprecht, T. Mietzner, *J. Med. Chem.* **41** (1998) 2553 (<https://doi.org/10.1021/jm970732a>)

12. A. Golbraikh, A. Tropsha, J. *Mol. Graph*. Model. **20** (2002) 269 (<https://doi.org/10.1016/S1093-3263(01)00123-1>)

13. L.F. Ramsey, W.D. Schafer, *The Statistical Sleuth. Wadsworth Publishing Company*, U.S.A. 1997 (<https://ir.library.oregonstate.edu/downloads/j3860c13r>)

14. A.J. Holder, D.M. Yourtee, D.A. White, A.G. Galaros, R.J. Smith Chain, *[J. Comput. Aided. Mol. Des](file:///C:\\Users\\Mounia-PC\\Downloads\\J.%20Comput.%20Aided.%20Mol.%20Des).* **17** (2003) 223 (<https://doi.org/10.1023/A:1025382226037>)

15. S. Chatterjee, A. S. Hadi, B. Price, Analysis of collinear data. In: Regression analysis by example, fourth ed. Wiley, New York, 2006, P. 221-258 (<https://doi.org/10.1002/0470055464.ch9>)

16. A. Tropsha, A. Golbraikh, *Curr. Pharm. Des.* **13** (2007) 3494 (<https://doi.org/10.2174/138161207782794257>)

17. W. Li, Y. Tang, Y.L. Zheng, Z.B. Qiu, *Bioorg. Med. Chem.* **14** (2006) 601 (<https://doi.org/10.1016/j.bmc.2005.08.052>)

18. R. Guha, D.T. Stanton, P.C. Jurs, *J. Chem. Inf. Model.* **45** (2005) 1109 (

<https://pubs.acs.org/doi/abs/10.1021/ci050110v>)

1. \*Corresponding author. E-mail: d\_messadi@yahoo.fr [↑](#footnote-ref-1)