Journal of the Serbian Chemical Society
JSCS-info@shd.org.rs • www.shd.org.rs/JSCS
J. Serb. Chem. Soc. *84* (12) 1405–1414 (2019) UDC 547.1+532.11+621.5.02:544.112:539.12
JSCS–5273 *Original scientific paper*

Quantitative structure–property relationship studies for the prediction of the vapor pressure of volatile organic compounds

MOUNIA ZINE[1], AMEL BOUAKKADIA[1,2], LEILA LOURICI[3] and DJELLOUL MESSADI[1]*

[1]*Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP. 12, 23000 Annaba, Algeria,* [2]*Abbes Laghrour University, Faculty of sciences and technology, Khenchela, Algeria and* [3]*Chadeli Ben Djedid University, BP. 73, 3600 El Taref, Algeria*

(Received 6 March, accepted 13 June 2019)

Abstract: A theoretical model (QSPR) using multiple linear regression analysis for predicting the vapor pressure (p_v) of volatile organic compounds (VOCs) has been developed. A series of 51 compounds were analyzed by multiple linear regression analysis. First, the data set was separated arbitrarily into a training set (39 chemicals) and a test set (12 chemicals) for statistical external validation. A four-dimensional model was developed using as independent variables theoretical descriptors derived from Dragon software when applying the GA (genetic algorithm)–VSS (variable subset selection) procedure. The obtained model was used to predict the vapor pressure of the test set compounds, and an agreement between experimental and predicted values was verified. This model, with high statistical significance (R^2 = 0.9090, Q^2_{LOO} = 0.8748, Q^2_{ext} = 0.8307, s = 0.24), could be used adequately for the prediction and description of the log p_v value of other VOCs. The applicability domain of MLR model was investigated using a William's plot to detect outliers and outsides compounds.

Keywords: molecular descriptors; VOCs; log p_v; multiple linear regression.

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 in the vapor state 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 from human activity.[1]

The sectors of activities that are more strongly transmitters of VOCs are road transport, industry, agriculture, and the tertiary sector. Other air pollutants that could be cited are biological contaminants (bacteria, pollen, fungi), physical con-

* Corresponding author. E-mail: d_messadi@yahoo.fr
https://doi.org/10.2298/JSC190306059Z

taminants (metals, particles, dust, radioactivity), and chemical contaminants that are part of the VOCs represented by gases (CO, O_3, NO_x, SO_2, fluorocarbons), dioxins and furans.[1]

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

As new compounds are emerging, other difficulties will also arise. Therefore, it is necessary to use 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, quantum chemical parameters of compounds can speedily and precisely obtained 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]

The vapor pressure of different compounds can be predicted from their formula and even unknown compounds can be identified 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 and determination of vapor pressure.

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

Different statistical techniques were used to develop the model to highlight the structural requirements for the ideal vapor pressure. The three objectives of the present paper were 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 the variation of log p_v value of these compounds.

EXPERIMENTAL

Data set

The p_v experimental values of 51 selected, structurally heterogeneous VOCs were collected from previous works,[4] and converted to log p_v 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 PM3method. All calculations were realized at the RHF (restricted Hartree–Fock) level with non configuration interaction. The molecular structures were optimized using the Polak–Ribiere algorithm and a gradient norm limit of 0.01 kcal $Å^{-1}$ mol^{-1}. The resulting 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 AssemblY).[6]

Descriptors with constant or near constant values inside each group were discarded. For each pair of correlated descriptors (with correlation coefficient $r \geq 0.95$), the one showing the highest pair correlation with the other descriptors was excluded. The genetic algorithm (GA)[7] was considered superior to other methods of the variable selection techniques. Thus, variable selection was performed on the training set using GA in the MobyDigs version of Todeschini[8] by maximizing the cross-validated explained variance Q^2_{LOO}.

Model development and validation

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

$$SDEC = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{1}$$

Cross validation techniques allow the assessment of the internal predictivity (Q^2_{LMO} cross validation; bootstrap) in addition to the robustness of the model (Q^2_{LOO} 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 the leave-one-out technique (LOO), otherwise the leave-more-out technique (LMO).

The LOO or LMO correlation coefficient, generally indicated with Q^2, was computed by evaluating the accuracy of the prediction of these "test" compounds:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS} \tag{2}$$

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^2 and the standard deviation error in prediction (*SDEP*).

$$SDEP = \sqrt{PRESS/n} \tag{3}$$

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 the LMO procedure (repeated 5000 times, with 4 objects left out at each step) was also performed ($Q^2_{L(4)O}$).

In the 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^2_{ext}, which is defined as:

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{n_{\text{ext}}} (\hat{y}_{i/i} - y_i)^2 / n_{\text{ext}}}{\sum_{i=1}^{n_{\text{tr}}} (y_i - \bar{y}_{\text{tr}})^2 / n_{\text{tr}}} = 1 - \frac{PRESS / n_{\text{ext}}}{TSS / n_{\text{tr}}} \tag{4}$$

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 from the training set, and external standard deviation error of the prediction ($SDEP_{ext}$), which is defined as:

$$SDEP_{\text{ext}} = \sqrt{\frac{1}{n_{\text{ext}}} \sum_{i=1}^{n_{\text{ext}}} \left(y_i - \bar{y}\right)^2} \tag{5}$$

where the sum runs over the test set objects (n_{ext}). According to Golbraikh and Tropsha,[12] a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{\text{CVext}} > 0.5 \tag{6}$$

$$R^2 > 0.6 \tag{7}$$

$$\left(r^2 - r^2_0\right) / r^2 < 0.1 \text{ or} \left(r^2 - r^{2\prime}_0\right) / r^2 < 0.1 \tag{8}$$

$$0.85 < K < 1.15 \text{ or } 0.85 < K' < 1.15 \tag{9}$$

Here:

$$r = \frac{\sum \left(y_i - \tilde{y}_i\right)\left(\tilde{y}_i - \bar{\tilde{y}}\right)}{\sqrt{\sum \left(y_i - \tilde{y}\right)^2 \sum \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2}} \tag{10}$$

$$r^2 = 1 - \frac{\sum (y_i - y_i^{r0})^2}{\sum (y_i - \bar{y})^2} \tag{11}$$

$$r'^2_0 = 1 - \frac{\sum \left(\tilde{y}_i - \tilde{y}_i^{r0}\right)^2}{\sum \left(\tilde{y}_i - \bar{\tilde{y}}\right)^2} \tag{12}$$

$$k' = \frac{\sum(y_i \tilde{y}_i)}{\sum(\tilde{y}_i)^2} \quad (13)$$

$$k = \frac{\sum(y_i \tilde{y}_i)}{\sum(y_i)^2} \quad (14)$$

where r is the correlation coefficient between the calculated and experimental values in the test set; r^2_0 (calculated *versus* observed values) and r_0^2 (observed *versus* calculated values) are the coefficients of determination; k and k' are the slopes of regression lines through the origin of the calculated *versus* observed and observed versus calculated, respectively y^{r0}_{i}, $\tilde{y}^{r0}_{\mathrm{i}}$; are defined as $y^{r0}_i = k\,\tilde{y}_{\mathrm{i}}$ and $\tilde{y}^{r0}_i = k' y_{\mathrm{i}}$ and the summations runs over the test set.

The applicability domain (AD) 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:

$$h_{\mathrm{i}} = x_i(X^TX)^{-1}x_i^T \; (i = 1,\ldots,n) \quad (15)$$

where x_i 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} of the compounds. The study we conducted consists of the multiple linear regressions (MLR) available in the MobyDygs software.

RESULTS AND DISCUSSION

In order to predict log p_{v}, 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:

$$\log p_{\mathrm{v}} = 11.0490 - 0.4602X0sol - 12.3322SpPosA\_H2 + \\ + 1.1372GATS2e - 1.2333Hy \quad (16)$$

Statistical parameters for the model with: $n_{\mathrm{tr}} = 39$ $n_{\mathrm{ext}} = 12$ are:

$$R^2 = 0.9009 \quad Q^2_{\mathrm{LOO}} = 0.8748 \quad Q^2_{\mathrm{Boot}} = 0.8555 \quad F = 77.25 \quad Q^2_{\mathrm{ext}} = 0.8307$$

$$SDEP = 0.256 \quad SDEC = 0.227 \quad SDEP_{\mathrm{ext}} = 0.297 \quad s = 0.24 \quad K_{\mathrm{xx}} = 38.51 \quad K_{\mathrm{xy}} = 45.57$$

The reported fitting and validation parameters have, as expected, high values indicating that the model has a 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).

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

Descriptor	Coefficient regressed	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

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] A smaller *t*-probability suggests a more significant descriptor. The *t*-probability values of the four descriptors are very small, indicating that all of them are highly significant descriptors.

The *VIF* 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* in Table I indicate a more serious multi-co linearity problem.

Models would not be accepted if they contain descriptors with *VIF*s above a value of five.[14]

$$VIF = \frac{1}{1 - R_j^2} \tag{17}$$

where R_j^2 is the squared correlation coefficient between the j^{th} coefficient regressed against all the other descriptors in the model.[15]

Applicability domain

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

A diagram of the statistical coefficients Q^2 and R^2 is presented Fig. 2 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 are smaller than those of the real models; the Q^2 values are lower (<10 %), as for the major part are the R^2 values (<30 %). This ensures that the established model has a real base, and is not due arbitrarity.

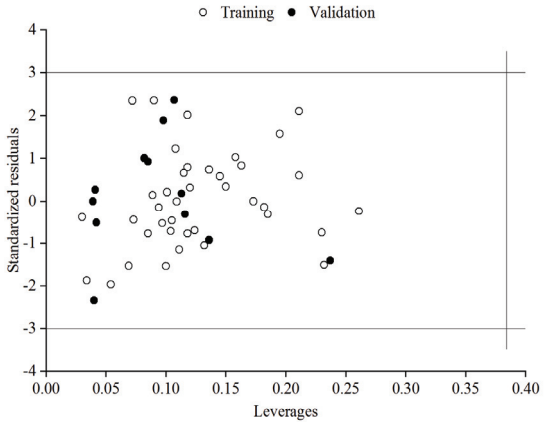

Fig. 1. Williams plot: jackknifed residuals and leverages.

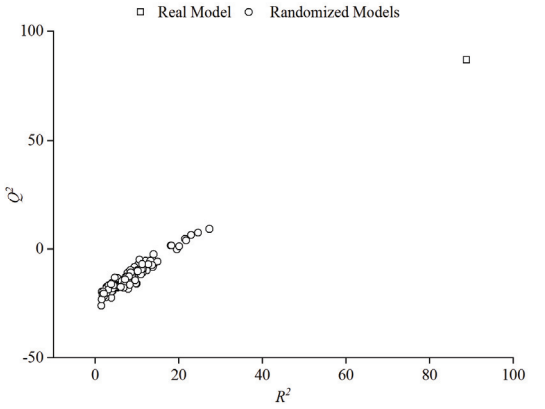

Fig. 2. Randomization test associated to the previous QSPR model.

This technique, plot of cross-validation log p_V *versus* experimental log p_V values, shown in Fig. 3, ensures the robustness of a QSPR model (see also Tables S-I and S-II of the Supplementary material to this paper).[16]

Descriptor contributions analysis

Based on a previously described procedure,[17,18] the relative contributions of the four descriptors to the model were 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 %).

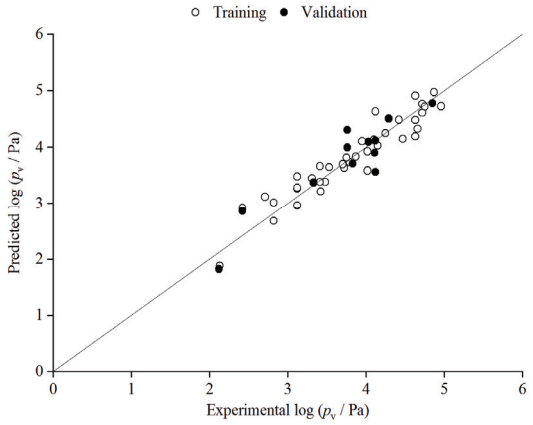

Fig. 3. Cross-validation *versus* experimental log p_v for the entire data set.

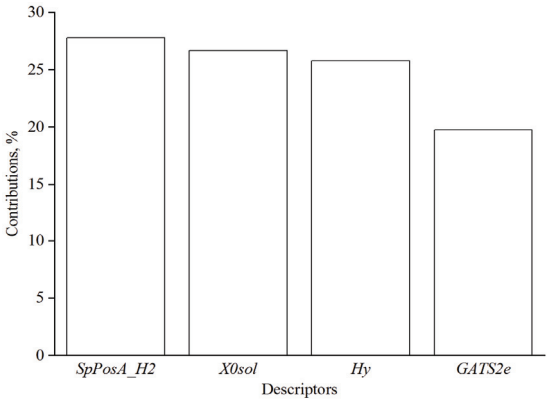

Fig. 4. Relative contributions of the selected descriptors to the MLR model.

CONCLUSIONS

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 were separated randomly into two subsets of 39 elements for training and 12 for external validation. The MLR model has good stability, 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 could be used successfully to estimate the vapor pressures for new compounds or for other compounds the experimental values of which are unknown.

SUPPLEMENTARY MATERIAL

Additional data are available electronically from http://www.shd.org.rs/JSCS/, or from the corresponding author on request.

И З В О Д

ПРОУЧАВАЊА КВАНТИТАТИВНИХ РЕЛАЦИЈА СТРУКТУРЕ И РЕАКТИВНОСТИ ЗА ПРЕДВИЂАЊЕ НАПОНА ПАРЕ ИСПАРЉИВИХ ОРГАНСКИХ ЈЕДИЊЕЊА

MOUNIA ZINE[1], AMEL BOUAKKADIA[1,2], LEILA LOURICI[3] и DJELLOUL MESSADI[1]

[1]Environmental and Food Safety Laboratory, Badji Mokhtar-Annaba University, BP. 12, 23000 Annaba, Algeria, [2]Abbes Laghrour University, Faculty of sciences and technology, Khenchela, Algeria и [3]Chadeli Ben djedid University, BP.73, 3600 El taref, Algeria

Развијен је теоријски модел (QSPR) коришћењем вишеструке линеарне регресионе анализе за предвиђање напона паре (p_v/Pa) испарљивих органских једињења (VOC), за серију од 51 једињења. На почетку је скуп података произвољно подељен у скуп за проучавање (39 једињења) и скуп за проверу (12 једињења) за екстерну статистичку валидацију. Развијен је четвородимензионални модел користећи као независне варијабле теоријске описнике изведене из софтвера Dragon применом GA (генетски алгоритам)–VSS (*variable subset selection*) процедуре. Добијени модел је употребљен за предвиђање напона паре једињења скупа за проверу, и потврђена је сагласност између експерименталних и претсказаних вредности. Овај модел, са високом статистичком значајношћу (R^2 = 0,9090, Q^2_{Loo} = 0,8748, Q^2_{Ext} = 0,8307, s = 0,24), може бити адекватно употребљен за предвиђање и описивање log (p_v/Pa) вредности других VOC. Област применљивости MLR модела испитивана је коришћењем Виљемсовог графа (*William's plot*) да се детектују једињења која се не уклапају у модел.

(Примљено 6. марта, прихваћено 13. јуна 2019)

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, R. Hmamouchi, M. Larif, M. Ghamali, M. Bouachrine, T. Lakhlifi, *J. Taibah Univ. Sci.* **10** (2016) 868 (https://doi.org/10.1016/j.jtusci.2015.04.007)
3. J. Akbar, S. Iqbal, F. Batool, A. Karim , 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*, 2nd ed., CRC Press Inc., Boca Raton, FL, 2006 (https://doi.org/10.1201/9781420044393)
5. *HyperChem 6.03 Package*, Hypercube, Inc., Gainesville, FL, 1999; software available at: http://www.hyper.com
6. *Talete Srl. Dragon for Windows (Software for Molecular Descriptor Calculation) Version 5.5*, Milano, 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, 2009 (http://www.talete.mi.it/)
9. L. Eriksson, J. Jaworska, A. Worth, M. McCronin, 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, Belmont, CA, 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.* **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*, 4th ed., S. Chatterjee, A. S. Hadi, Eds., Wiley, New York, 2006, pp.. 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).