EQA - International Journal of Environmental Quality
ISSN 2281-4485 - Vol. 52 (2022): 20-28
Journal homepage: https://eqa.unibo.it/

published by

QSAR model for pk_a prediction of phenols

Hakim Hamada

Materials and Environment Analytical Sciences Laboratory, University of Oum El Bouaghi, Algeria

*Corresponding author E-mail: hakim.hamada@univ-oeb.dz

Article info
Received 28/10/2022; received in revised form 10/12/2022; accepted 30/12/2022
DOI: 10.6092/issn.2281-4485/15686

Abstract

Descriptors (topological, mathematical and quantum) were used to generate quantitative construction property connections (QSPR) for the pKa of 80 phenols. The informational index was divided into 56 preparation and 24 test sets, and models were built using the preparation set's incomplete least squares (PLS) relapse. The consistency and predictive power of the best acquired QSAR models were achieved through internal approval, Y randomization, and external approval, and their pertinence area was confirmed by the influence technique. The benefits of the various direct relapse investigations' measurable boundaries. Standard deviation (S), standard deviation error of prediction (SDEP, External validation coefficient test), determination coefficient R^2, cross-validated R^2 (Q^2) (SDEPext). The cross-validated R^2 (test Q^2ext) values (95.68%, 95.22%, 0.304, 0.312, 0.292, and 96.24%, respectively) attest to the model's good fit.

Keywords
randomization; descriptors; regression; statistical; parameters; stability; external validation

Introduction

A fundamental physical property of drugs is the acid dissociation constant, which describes the ionization of compounds in aqueous solution. When comparing the properties of neutral and ionized species, it is common to find differences in solubility, and thus absorption, bioreactivity, and toxicity. Furthermore, for financial reasons, analysts work on developing strategies to anticipate, which can be less tedious, more financial, and simple. One of the primary options is to use a quantitative construction natural action/property relationship (Dearden, 2016; McKinney et al., 2000). which includes numerically determined decisions that quantitatively depict movement and property regarding atomic characteristics, such as descript-pinnacles of compound designs created with PC-based innovation (ROY et al., 2015). The method of quantitative relationships with structure of activity was used to predict the calculation of the ionization constant for a group of phenols.

Material and Methods

A Quantitative Structure-Activity Relationshio (QSAR) modeling was developed to study the constant acidity of a series of phenols using descriptors calculated by Dragon version 5.3 (Todeschini et al., 2006) and hyper hem 7.5 software (HYPERCHEM™ RELEASE 7.2000). The genetic algorithm (LEARD et al., 1992) is thought to be superior to other variable determination techniques. As a result, factor determination on the preparation set was performed, involving genetic algorithm (GA) in Todeschini's adaptation of Moby Digs (TODESCHINI et al., 2009) by maximizing the variance explained by cross-validation by omitting an observation. Ordinary least squares regression and genetic algorithm selection of explanatory variable subsets (GA -VSS) (Pavan et al., 2004). The crossover and mutation processes of the genetic algorithm in the Moby Digs software are controlled by a parameter T ranging from 0 to 1. The genetic algorithm's parameters were set as follows: Pop = 100 for the model population; T = 0.5 to balance the roles of the two processes of crossover and mutation.

The use of the GA-VSS method has resulted in several good models for predicting acid dissociation constant at logarithmic scale (pka=-logka) based on various sets of molecular descriptors. The The mean atomic van der Waals volume(Mv) and mean electrotopological state (Ms) were used to create the best model.

Experimental data

Compounds evaluated are listed in Table 1.

Acidity ka *(pka=-logka)* (Pirsellova et al.,1998).

The experimental values of phenol dissociation constants were discovered in the literature. *Mv* denotes the mean atomic van der Waals volume (scaled on the carbon atom). Divide the sum of the van der Waals volumes by the number of atoms to get the mean volume (Mv).

$$Mv = \frac{Sv}{nAT} \qquad [1]$$

where *nAT* denotes the number of atoms and *Sv* denotes the sum of the van der Waals volumes

$$Sv = \sum_{i=1}^{A} V_i \qquad [2]$$

mean electrotopological condition (Ms) (Todeschini et al., 2000). Divide Ss by the number of non-hydrogen

Table 1. *Data matrix of independent variables (molecular descriptors) and dependent variables*

Object	N°	Status	Y Exp.	PKA	Ms	Mv
2,3,4,5-tetrachlorophenol	1	Training	6.22	6.22	2.98	0.85
2,3,5 -trimethylpheno	2	Training	10.48	10.48	2.27	0.6
2,3,5,6-tetrafluorophenol	3	Training	5.99	5.99	4.39	0.67
2,3-dimethylphenol	4	Training	10.34	10.34	2.33	0.61
2,4-dibromophenol	5	Training	7.87	7.87	2.5	0.81
2,4-dimethylphenol	6	Training	10.52	10.52	2.33	0.61
2,6-difluorophenol	7	Training	7.51	7.51	3.67	0.66
2,6-diphenylphenol	8	Training	9.92	9.92	2.12	0.69
2-acetylphenol	9	Training	9.19	9.19	2.8	0.63
2-allylphenol	10	Training	9.92	9.92	2.38	0.63
2-bromo-4-methylphenol	11	Training	8.67	8.67	2.42	0.69
2-chlorophenol	12	Training	8.55	8.55	2.68	0.69
2-ethylphenol	13	Training	10.2	10.2	2.31	0.61
2hydroxy benzaldhyde	14	Training	8.34	8.34	2.93	0.65
2-hydroxybenzamide	15	Training	8.36	8.36	3.00	0.64
2-hydroxybenzylalcohol	16	Training	9.92	9.92	2.76	0.61
2-isopropylphenol	17	Training	10.4	10.4	2.27	0.6
2-methylphenol	18	Training	10.26	10.26	2.42	0.62
2-tert-butylphenol	19	Training	10.62	10.62	2.23	0.59
3 -methoxyphenol	20	Training	9.65	9.65	2.54	0.61
3,4,5,6-tetrabromo-2-methylphenol	21	Training	6.42	6.42	2.42	0.89
3,4,5-trimethylpheno	22	Training	10.5	10.5	2.27	0.6
3,5-dichlorophenol	23	Training	8.18	8.18	2.80	0.75
3-acetylphenol	24	Training	9.19	9.19	2.80	0.63
3-chlorophenol	25	Training	9.1	9.1	2.68	0.69
3-cyanophenol	26	Training	8.61	8.61	2.87	0.69
3-ethylphenol	27	Training	10.07	10.07	2.31	0.61
3-hydroxybenzaldehyde	28	Training	9	9	2.93	0.65

Table 1 (continued). *Data matrix of independent variables (molecular descriptors) and dependent variables*

Object	N°	Status	Y Exp.	PKA	Ms	Mv
3-iodophenol	29	Training	8.88	8.88	2.43	0.75
3-methylphenol	30	Training	10	10	2.42	0.62
3-phenylphenol	31	Training	9.63	9.63	2.23	0.67
3-tert-butylphenol	32	Training	10.12	10.12	2.23	0.59
4-bromo-2,6-dimethylpheno	33	Training	10.01	10.01	2.34	0.66
4-bromo-6-chloro-2-methylphenol	34	Training	8.2	8.2	2.55	0.73
4-chloro-2-iso-propyl-5-methylphenol	35	Training	10.03	10.03	2.34	0.62
4-chloro-3 -methylphenol	36	Training	9.55	9.55	2.57	0.66
4-chloro-3,5-dimethylpheno	37	Training	9.7	9.7	2.48	0.64
4-chloro-3,5-dimethylphenol	38	Training	9.7	9.7	2.48	0.64
4-heptyloxyphenol	39	Training	10.7	10.7	2.12	0.57
4-hexyloxyphenol	40	Training	10.7	10.7	2.17	0.58
4-hromo-2,6-dichlorophenol	41	Training	6.4	6.40	2.76	0.83
4-hydroxyazobenzene	42	Training	8.78	8.78	2.33	0.68
4-hydroxybenzamide	43	Training	9.23	9.23	3.00	0.64
4-hydroxybenzophenone	44	Training	8.89	8.89	2.51	0.68
4-iso-propylphenol	45	Training	10.30	10.3	2.27	0.60
4-methoxyphenol	46	Training	10.20	10.2	2.54	0.61
4-methylphenol	47	Training	10.26	10.26	2.42	0.62
4-propylphenol	48	Training	10.3	10.3	2.23	0.60
4-sec-butylphenol	49	Training	10.3	10.3	2.20	0.59
4-tert-butylphenol	50	Training	10.23	10.23	2.23	0.59
4-tert-pentylphenol	51	Training	10.3	10.3	2.17	0.58
ethyl-3-hydroxybenzoate	52	Training	9.09	9.09	2.75	0.61
methyl-4-hydroxybenzoate	53	Training	9.05	9.05	2.86	0.63
pentabromophenol	54	Training	4.57	4.57	2.48	1.06
pentachlorophenol	55	Training	5.1	5.10	3.05	0.91
phenol	56	Training	9.99	9.99	2.52	0.64
2,3,5-trichloropheno	57	Test	6.75	6.75	2.90	0.80
2,3,6-trimethylphenol	58	Test	10.63	10.63	2.27	0.60
2,3-dichlorophenol	59	Test	7.58	7.58	2.80	0.75
2,4,5 -trichloropheno	60	Test	7.37	7.37	2.90	0.80
2,4,6-tribromophenol	61	Test	6.31	6.31	2.49	0.89
2,4-dichlorophenol	62	Test	7.87	7.87	2.80	0.75
2,5-dichlorophenol	63	Test	7.58	7.58	2.80	0.75
2,5-dimethylphenol	64	Test	10.34	10.34	2.33	0.61
2-bromophenol	65	Test	8.45	8.45	2.51	0.72
2-fluorophenol	66	Test	8.73	8.73	3.17	0.65
2-phenylphenol	67	Test	9.55	9.55	2.23	0.67
2-tert-butyl-4-methylphenol	68	Test	11.39	11.39	2.19	0.58
3,4-dimethylphenol	69	Test	10.32	10.32	2.33	0.61
3,5-dimethylphenol	70	Test	10.15	10.15	2.33	0.61
3-iso-propylphenol	71	Test	10.1	10.1	2.27	0.60
4-butoxyphenol	72	Test	10.6	10.6	2.35	0.59
4-chloro-2-methylphenol	73	Test	9.67	9.67	2.57	0.66
4-cyanophenol	74	Test	7.96	7.96	2.87	0.69
4-ethoxyphenol	75	Test	10.5	10.5	2.43	0.60
4-ethylphenol	76	Test	10	10	2.31	0.61
4-hydroxybenzylcyanide	77	Test	9.52	9.52	2.73	0.66
4-hydroxyphenethylalcohol	78	Test	9.92	9.92	2.63	0.60
methyl-3-hydroxybenzoate	79	Test	9.21	9.21	2.86	0.63
pentafluorophenol	80	Test	5.86	5.86	4.67	0.68

(nSK) to get the mean electrotopological state (Ms):

$$Ms = \frac{Ss}{nSK} \qquad [3]$$

Electro topological (Ss) Kier Hall states

$$Ss = \sum_{i=1}^{A} S_i \qquad [4]$$

Results and Discussion

Cross-validated. Determination coefficient calculation formula (R^2)-Chatterje et al., 2006). R^2 The quality of fit is indicated by the coefficient of determination (R^2), which is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \qquad [5]$$

where: Yi= ith observed response value;
$\hat{y}_i$ = ith fitted response; $\bar{y}_i$ = mean response;

R^2 equals 1 for the ideal model, where the sum of squared residuals is 0.

The model's ability to fit data worsens as R^2's value deviates from 1. R^2 is the multiple correlation coefficient and its square root (R).

Adjusted determination coefficient (Adjusted R^2) (Besse, 2003).

$$R_a^2 = 1 - \frac{SCE/_{n-k}}{SCT/_{n-1}} = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) \qquad [6]$$

where: n refers to the number of observation and k number of descriptors.

If the number of descriptors in a model is increased for a fixed number of observations, R^2 values will always increase, but this will result in a decrease in degree of freedom and low statistical reliability. As a result, a high R^2 value does not always indicate a good statistical model that fits the available data well.

Ratio (The Fisher-Snedécor Coefficient) (Bando et al., 1994).

The statistical significance of the regression equation at specified degrees of freedom is determined by the F statistic, which is calculated from R2 and the number of data points (df)

The F-ratio test is one of the most well-known statistical tests, and it is defined as follows:

$$F = \frac{n-1-k}{k} \times \frac{R^2}{1-R^2} \qquad [7]$$

Standard deviation (s) (Siegel, 1997).

$$s = \sqrt{\frac{\sum_1^n \left(y_{i;\,exp} - y_{i;\,calc}\right)^2}{n-1-k}} = \sqrt{\frac{RSS}{n-1-k}} \qquad [8]$$

It is a metric for dispersion. It describes how data distribution is done around the average. The closer it is to zero, the more accurate the adjustment and prediction.

The Predicted residual sum of squares (Press and Wilson, 1978; Hastie et al., 2009).

The most important parameter for estimating the models' true predictive error appears to be the PRESS (Predicted Residual Sum of Squares) statistic. Its low value implies that the model outperforms chance and is statistically significant. It is calculated using the following equation:

$$PRESS = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2 \qquad [9]$$

Residuals (e_i). The disparity between observed and predicted or fitted values. The fitted model does not account for this aspect of the observation. An observation's residual is: yi=ith observed response value; $\hat{y}_i$= ith fitted response.

Leverages(hii) (Press et al,. 1978). Indicate whether the observed predictor values are unusual in relation to the rest of the data. High leverage observations may have a significant impact on the fitted value, and thus the regression model.

Leverages are obtained from the hat matrix (H), which is a n x n projection matrix specified as:

$$Hii = X_i^t(X^TX)^{-1}X_i \qquad [10]$$

The ith diagonal element, h_i of H, is the leverage of the ith observation. If h_i is large, the ith observation has out-of-the-ordinary predictors (x_{1i} .x_{2i}.....x_{ki}), that is, predictor values that are far from the mean

DOI: 10.6092/issn.2281-4485/15686

vector $(\overline{X}_1.\overline{X}_2....\overline{X}_K$) utilizing the Mahalanobis distance: n= number of observations, k= number of predictors, Xki=ith observation of the kth predictor, $\overline{X}_{ki}$ = ith predictor mean, X= response matrix, Y= predictor matrix. Standard deviation error in calculation defined as (Golbraikh et al., 2002).

$$\mathrm{SDEC} = \sqrt{\frac{\Sigma_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \qquad [11]$$

Standard deviation error of prediction (SDEP) (Roy et al., 2015).

$$\mathrm{SDEP} = \sqrt{\frac{\mathrm{PRESS}}{n}} \qquad [12]$$

Cross-validated R^2 (R^{2}cv) (or Q^2) (Consonniv et al., 2012),

$$Q^2 = 1 - \frac{\Sigma_1^n\left(y_{\text{obs; train}} - y_{\text{pred; train}}\right)^2}{\Sigma_1^n\left(y_{\text{obs; train}} - \bar{y}_{\text{train}}\right)^2} \qquad [13]$$

A value Q2 > 0.5 is considered good, and a value Q2 > 0.9 is considered excellent. Q2ext is the external validation coefficient (Golbraikh et al., 2002).

$$Q^2_{\text{ext}} = 1 - \frac{\Sigma_{\&}^{\text{next}}\left(\hat{y}_{i/i} - y_i\right)^2}{\Sigma_1^{\text{next}}\left(y_{\text{ops; train}} - \bar{y}_{\text{trai}}\right)^2 \big/ \text{ntrain}} = \frac{\text{PRESS}/\text{next}}{\text{TSS}/\text{ntrain}} \qquad [14]$$

here, n$_{ext}$ refers to the number of test set compounds.

$$\mathrm{SDEP}_{\text{ext}} = \sqrt{\frac{1}{\text{next}}\Sigma_1^{\text{next}}\left(y_{i;\exp} - y_{i;\text{pred}}\right)^2} \qquad [15]$$

where the sum is applied to the test set objects (ext n).

QSAR Model development and validation.

The optimal model's equation is as follows:

pKa = 21.216-1.586Ms-12.022Mv [16]

Statistical parameters for the model is presented in table 2

Table 2
Statistical parameters

//ntr	**next**	**S**	**Q² %**	**R² %**	**F**
56	24	0.304	95.22	93.64	586.64
R²adj %	Q² ext %	SDEC	SDEP	SDEP ext	P
95.51	96.24	0.296	0.312	0.292	00

Where: ntr = the number of training set); *n*ext :the number of objects in the external set; S = Standard deviation; F = the Fisher-Snedécor Coefficient; R^2 and Q^2 values attest to the model's good fitting performance, which is also very significant (great value of the Fisher parameter F). The model is robust, with only a 3% difference between R^{2}and Q^2.
A value Q^2 greater than 0.5 is generally considered good, and a good standard error s = 0.304 and P less than 0.05 indicates that the regression equation has statistically significant statistical parameters.
SDEPex= SDEP ; performs better in external prediction than in internal prediction Q$^2_{ex}$ > Q^2 criteria of Tropsha et al. (2003).

According to the literature (Golbraikh et al., 2002), an additional external validation is applied solely to the test set. A predictive QSPR model must meet the following conditions, according to Tropsha et al. (2003) recommended .'s criteria:

$$Q^2_{ext} > 0.6 \qquad [17]$$

$$R^2_{ext} > 0.7 \qquad [18]$$

$$\frac{\left|R^2 - R_0^2\right|}{R^2} < 0.3 and 0.85 \le K \le 1.15 \qquad [19]$$

DOI: 10.6092/issn.2281-4485/15686

$$\frac{\left|R^2 - R_0'^2\right|}{R^2} < 0.3 and 0.85 \le K' \le 1.15 \qquad [20]$$

$$\left|R_0^2 - R_0'^2\right| < 0.3 \qquad [21]$$

$$K' = \frac{\sum_1^n \left(y_i \hat{y}_i\right)^2}{\sum_1^n \left(y_i\right)^2} \qquad [22]$$

$$K = \frac{\sum_1^n \left(y_i \hat{y}_i\right)^2}{\sum_1^n \left(\hat{y}_i\right)^2} \qquad [23]$$

$$R_O^2 = 1 - \frac{\sum_1^n \left(\hat{y}_i - y_i^{r0}\right)^2}{\sum_1^n \left(y_i - \bar{y}\right)^2} \qquad [24]$$

$$R_O'^2 = 1 - \frac{\sum_1^n \left(y_i - \hat{y}_i^{r0}\right)^2}{\sum_1^n \left(y_i - \bar{y}\right)^2} \qquad [25]$$

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 slopes of regression lines through the origin of the calculated versus observed and the observed versus calculated, respectively:

$$y_i^{r0} = k\hat{y} \qquad [26]$$

and

$$\hat{y}_i^{r0} = k'y \qquad [27]$$

The sums are calculated over all samples in the test set. The reason for using R_0^2 and requiring k values close to 1 is that when comparing actual versus predicted properties, an exact fit is required, not just a correlation. An example of a regression between observed vs. predicted (Fig. 1) and predicted vs. observed (Fig. 2) activities for compounds from an external testset.

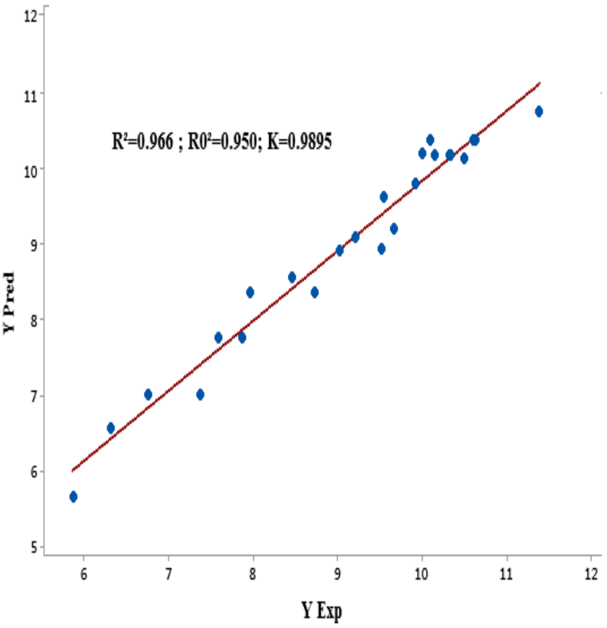

Figure 1. *Plot of experimental vs. predicted values in a regression model*

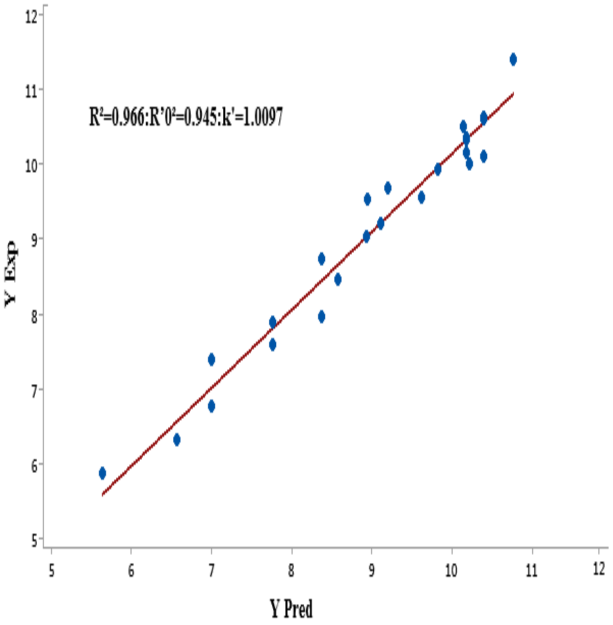

Figure 2. *Plot predicted of vs. experimental values in a regression model*

DOI: 10.6092/issn.2281-4485/15686

$$Q^2_{ext} = 0.924 > 0.6 \qquad [28]$$

$$R^2 = 0.966 > 0.7 \qquad [29]$$

$$\frac{\left|R^2 - R_0^2\right|}{R^2} = \frac{\left|0.966 - 0.950\right|}{0.966} = 0.016 < 0.3 \text{and} 0.85 \le K = 0.9895 \le 1.15 \qquad [30]$$

$$\frac{\left|R^2 - R_0'^2\right|}{R^2} = \frac{\left|0.966 - 0.945\right|}{0.966} = 0.022 < 0.3 \text{and} 0.85 \le K' = 1.0097 \le 1.15 \qquad [31]$$

$$\left|0.950 - 0.945 = 0.005\right| < 0.3 \qquad [32]$$

Applicability domain of the model (Eriksoon et al., 2003; Tropsha et al., 2003).
A common definition of the AD is based on the following leverage values

$$h_{ii} = X_i^t\left(X^tX\right)^{-1}X_i \qquad [33]$$

i=1,2,3………n " or each compound, where i is the query compound's descriptor row-vector and X is the matrix of k model descriptor values for n training set compounds. Compounds with h > h* (h* being a threshold value equal to 3p/n, where p is the number of descriptors in the model plus one and n is the number of compounds in the training set) are chemically distinct from the training set compounds and thus outside the AD.
He determined warning leverage (h* = 0.161).
Figure 3 depicts the end result. Chemicals, with the exception of compounds No. 3, 15, 77, and 79, are within the AD. Even for chemicals with h values greater than h*=0.161, the predicted pka values are close to the experimental values. Furthermore, all compounds fall within the standard residuals of ±2 (s.d).. As a result, the developed models for these sets are also reasonable.

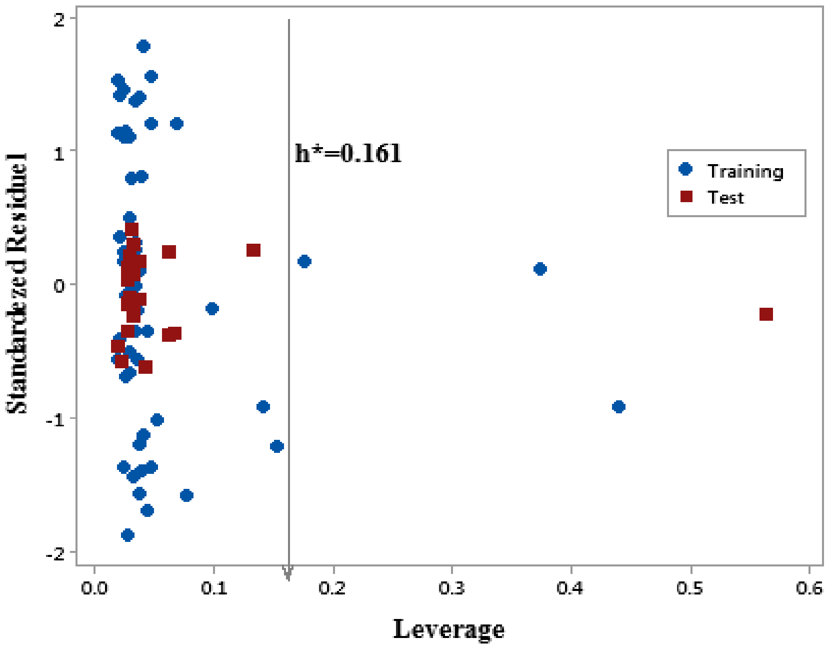

Figure 3
The Williams plot of the leverage value against the standardized residuals

DOI: 10.6092/issn.2281-4485/15686

Randomization test (Tropsha and Golbraikh, 2007). This is a common technique for ensuring the robustness of a QSAR model. The dependent-variable vector, Y-vector, is randomly shuffled in this test, and a new QSAR model is created using the original independent-variable matrix.
The low Q² and R². The values obtained after each shuffle show (Table3) that the good results in our original model are not due to a random correlation of the training set.

Table3 *Randomization test*

Iteration	R² %	Q² %
0	95.68	95.22
1	2.67	00
2	1.99	00
3	4.9	00
4	6.44	00
5	6.19	00
6	6.4	00
7	0.39	00
8	0.13	00
9	0.2	00
10	2.57	00

Conclusions

The QSAR method was used to predict the acidity of phenols. We discovered two critical descriptors that accurately predict the pka (Mw and Me).
The models' validity has been established through the selection of appropriate statistical parameters. We found a strong correlation between experimental and predicted activity values, indicating that the QSAR model was validated and of high quality.

References

BANDO P., MARTIN N., SEGURA J.L., SEOANE C., ORTI E., VIRUELA P.M., CANO F.H. (1994) Single-Component Donor-Acceptor Organic Semiconductors Derived from TCNQ. The Journal of Organic Chemistry, 59(16):4618–4629. https://doi.org/10.1021/jo00095a042

BESSE P.(2003) Pratique de la modélisation statistique ; Publication du laboratoire de statistique et Probabilité ,P:11 http://www.math.univ-toulouse.fr/~besse/ pub/modlin.pdf

CHATTERJE S., HADI A.S. (2006). Regression Analysis by Example. 4th Edition, John Wiley & Son, Inc., Hoboken, PP :366. ISBN: 978-0-470-05545-8

CONSONNI V., BALLABIO D., TODESCINI R. (2010) Evaluation of model predictive ability by external validation techniques. Journal of Chemometrics, 24(3-4): 194–201. https://doi.org/10.1002/cem.1290.

DEARDEN J.C. (2016) The History and Development of Quantitative Structure-Activity Relationships (QSARs). International Journal of Quantitative Structure-Property Relationships, 1(1):1–44. https://doi.org/10.4018/IJQSPR.2016010101

ERIKSSON L., JAWORSKA J., WORTH A.P., CRONIN M.T.D., McDOWELL, R.M., GRAMATICA P. (2003) Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs. Environmental Health Perspectives, 111(10): 1361–1375. https://doi.org/ 10.1289/ehp.5758.

HASTIE T., TIBSHIRANI R ., FRIEDMAN J H. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. (Springer Series in Statistics), pp:307 ISBN: 0387848576

HYPERCHEM™ RELEASE 7 (2000) Hypercube for Windows, Molecular Modeling System, http://www.hyper.com

LEARDI R., BOGGIA R., TERRILE M. (1992) Genetic algorithms as a strategy for feature selection. Journal of Chemometrics, 6(5):267–281. https://doi.org/10.1002/cem.1180060506

MCKENNEY J.D., RICHARD A., WALLER C., NEWMAN M.C., GERBERICK F. (2000) The practice of structure activity relationships (SAR) in toxicology. Toxicological Sciences, 56(1):8-17. https://doi.org/10.1093/toxsci/56.1.8

PAVAN, M., MAURI, A., & TODESCHINI, R. (2004). Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority settings. Analytical and Bioanalytical Chemistry, 380(3), 430–444. https://doi.org/10.1007/s00216-004-2762-3

PIRŠELOVÁ K.; BALÁŽ Š.; SCHULTZ T.W. (1996) Model-Based QSAR for Ionizable Compounds: Toxicity of Phenols Against Tetrahymena pyriformis. Archives of Environmental Contamination and Toxicology, 30:170–177. http://dx.doi.org/10.1007/BF00215795

DOI: 10.6092/issn.2281-4485/15686

PRESS, S. J., & WILSON, S. (1978) Choosing between Logistic Regression and Discriminant Analysis. Journal of the American Statistical Association, 73(364):699–705. http://dx.doi.org/10.1080/01621459.1978. 10480080

ROY,K, S. KAR, R.N. DAS, (2015) Understanding the Basics of QSAR forApplications in Pharmaceutical Sciences and Risk Assessment, Academic Press. 1st Edition , pp. 254-258 ISBN 9780128016336

TODESCHINI R., BALLABIO D., CONSONNI V., MAURI A., PAVAN M. (2009) MOBYDIGS, version 1.1, Copyright TALETE srl.2004, http://www.disat.unimib.it
.

TODESCHINI R., CONSONNI V. (2000) Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry. https://doi.org/10.1002/9783527613106

TODESCHINI R., CONSONNI V. PAVAN V. (2006) DRAGON Software for the Calculation of Molecular Descriptors, Release 5.4 for Windows, Milano http://www.disat.unimib.it