CARPATHIAN J. MATH.
Volume **40** (2024), No. 3,
Pages 607-622

Online version at `https://www.carpathian.cunbm.utcluj.ro/`
Print Edition: ISSN 1584 - 2851; Online Edition: ISSN 1843 - 4401
DOI: https://doi.org/10.37193/CJM.2024.03.04

Robust mathematical models of structure and quantitative structure-property relationship studies of alkyl halides

ZOIŢA BERINDE and CLAUDIA BUTEAN

ABSTRACT. A QSPR (**Q**uantitative **S**tructure-**P**roperty **R**elationship) model links mathematically various physicochemical properties with the structure of a molecule. Establishing such quantitative relationships is of great technological importance as in this way one can predict the properties of new untested molecules by means of a linear or nonlinear equation that expresses a certain property as an explicit function of one or more independent variables. These functional relationships (usually called QSPR models) are obtained by performing specific QSPR studies on a class of similar compounds whose properties have been already determined and were appropriately correlated to their molecular structures.

In this study we are interested to determine the best QSPR models for the normal boiling point (*bp*) and molar refraction (mr) of 64 alkyl halides by using the topological index ZEP, the structural parameter (H_d) and the number of carbon atoms (N) as predictor variables. In order to validate the developed models with respect to the goodness-of-fit, robustness and predictive ability for the two properties considered in the class of alkyl halides, external validation, cross-validation (leave-one-out) and randomization (y-randomization) were performed. The obtained results have shown that the three descriptors above could be efficiently used for modelling and predicting the normal boiling points and molar refractions of the considered class of compounds.

1. INTRODUCTION

By a QSPR (Quantitative Structure-Property Relationship) study it is possible to link mathematically various physicochemical properties with the structure of a molecule. Similarly, a QSAR (Quantitative Structure-Activity Relationship) enables scientists to determine a mathematical expression that optimally links a biological activity with the structure of a chemical compound, see Kier and Hall [30], Trinajstić [46], [47], Balaban et al. [2], [3], Diudea and Ivanciuc [21], Berinde [7], Diudea et al. [22], Engel and Gasteiger [24]. From the technological point of view it is of great importance to find such quantitative relationships (QSPR, QSAR) as in this way one can predict the properties of new untested molecules by means of a linear or nonlinear functional expression that involves one or more variables.

Such a functional relationship (usually called a *QSPR / QSAR model*) is obtained by performing specific QSPR / QSAR studies on a class of similar compounds whose properties have been already determined and then correlated to their molecular structures.

The mathematical modelling of the relationship between the structure of a chemical compound and its physicochemical properties or biological activities represents the main object of molecular topology, constituted as an interdisciplinary branch of graph theory, applied in the study of chemical structures, with important applications in chemistry, biochemistry, and in the pharmaceutical industry.

This mathematical modelling process goes through three stages:

(1) Modelling the chemical structure with the help of molecular graphs;

Received: 22.09.2023. In revised form: 02.02.2024. Accepted: 12.02.2024

2010 *Mathematics Subject Classification.* 92E10, 05C90.

Key words and phrases. *topological index, multiple linear regression, quantitative structure-property relationship (QSPR), alkyl halides.*

Corresponding author: Zoiţa Berinde; zoita.berinde@cb.utcluj.ro

(2) Representation of molecular graphs by a matrix, a number, a string of numbers or a polynomial etc., that allows to obtain topological descriptors/indices (TI);
(3) Using topological indices to find quantitative mathematical relationships structure-property (QSPR) or biological structure-activity (QSAR).

The most used mathematical tools for constructing QSPR /QSAR models are the Linear Regression Method and the Least Squares Method. There is a rich literature on both QSPR and QSAR studies for various properties / activities and by using several topological indices as independent variables.

For example, some linear and non-linear simple regression models were proposed in Berinde [8] for the boiling point and molar refraction in a small set (16-18 compounds) of alkyl halides, by using a single structural descriptor, the ZEP index. For the boiling point (bp), the following six QSPR equations, with the best value of the correlation coefficient $r = 0.977$, were obtained in Berinde [8]:

$$bp = -101.32922 + 9.8283528 \cdot ZEP;\ r = 0.950;\ s = 12.02\,(N = 18)$$

$$bp = -144.22333 + 17.938117 \cdot ZEP^{0.5} * \ln ZEP;\ r = 0.951;\ s = 11.96\,(N = 18)$$

$$bp = -111.74676 + 10.264089 \cdot ZEP;\ r = 0.967;\ s = 10.17\,(N = 17)$$

$$bp = -112.88149 + 10.253969 \cdot ZEP;\ r = 0.977;\ s = 8.7\,(N = 16)$$

$$bp = \left(26.217166 - 109.79589 \cdot \frac{\ln ZEP}{ZEP}\right)^2;\ r = 0.968;\ s = 9.9\,(N = 17)$$

$$\ln(bp) = 6.0654915 - 135.92523 \cdot ZEP^{-1.5};\ r = 0.977;\ s = 8.5\,(N = 18)$$

For the molar refraction (mr), five QSPR equations were also obtained in [8], with the best value of the correlation coefficient $r = 0.98$:

$$mr = -2.3500175 + 1.5734065 \cdot ZEP;\ r = 0.970;\ s = 1.8\,(N = 17)$$

$$mr = 75.493286 - 22479.109 \cdot \frac{\ln ZEP}{ZEP^2} + \frac{48761.419}{ZEP^2};\ r = 0.970;\ s = 1.7\,(N = 17)$$

$$mr = 5.0938819 - 11.422281 \cdot \frac{\ln ZEP}{ZEP};\ r = 0.970;\ s = 1.7\,(N = 17)$$

$$mr = -2.8525349 + 1.5850079 \cdot ZEP;\ r = 0.980;\ s = 1.5\,(N = 17)$$

$$\frac{1}{mr} = 0.010392272 + 3.180691 \cdot \frac{\ln ZEP}{ZEP^2};\ r = 0.980;\ s = 1.51\,(N = 16).$$

Several QSPR studies for the normal boiling points were also performed for various classes of compounds:

- Balaban et al. [2] correlated the normal boiling points at normal pressure of 532 halogenated alkanes C1-C4 with various molecular connectivity indices due to Balaban, Randić and Kier and Hall, see Berinde [15] for their definition;
- Duchowicz et al. [23] used DRAGON5 evaluation software to select the best molecular descriptors from a set of more than thousand of rigid molecular descriptors in order to predict the normal boiling points of 200 organic molecules;
- Gharagheizi et al. [27] proposed a comprehensive, reliable, and predictive model using a large dataset of pure chemical compounds;
- Arjmand and Shafiei [1] predicted the normal boiling points and enthalpy of vaporizations of alcohols and phenols;
- Dai et al. [20] predicted separately the normal boiling points for 80 alkanes, 65 unsaturated hydrocarbons and 70 alcohols;

- Sola et al. [44] developed QSPR models for the prediction of the normal boiling point of organic compounds (and also of the critical temperature and the critical pressure);
- for other related contributions for various classes of compounds, see also Balaban et al. [3], Basak et al. [5], Carlton [18], Ivanciuc et al. [29], Öberg [36], Sanghvi and Yalkowsky [43], Wei [49] and some of the papers cited there.

Except for the paper by Berinde [8], it appears that no other study has dealt with the prediction of boiling points and / or molar refractivity of alkyl halides. Instead, some other properties were studied for this class of compounds or for related ones:

- Xu et al [50] obtained QSPR models for sub-cooled liquid vapor pressures (lg pL), n-octanol/water partition coefficient (lg Kow) and aqueous solubilities (lg Sw, L) of halogenated anisoles;
- Gajewicz et al. [25] derived QSPR models for the logarithmic values of the sub-cooled liquid vapor pressure (log pL) for a large set of polychlorinated and polybrominated congeners of benzenes, biphenyls, dibenzo-p-dioxins, dibenzofurans, diphenyl ethers and naphthalenes;
- Lu et al. [34] proposed a robust model for estimating thermal conductivity of liquid alkyl halides;

In the present paper we are interested to fill this gap and determine the best QSPS models for the normal boiling point (bp) and molar refraction (mr) for a significantly larger class of 64 alkyl halides, by using as independent variables the topological index ZEP, the structural parameter (H_d) and the number of carbon atoms (N) as predictor variables. To derive the models we use the simple and multiple linear regression.

We have considered a class of alkyl halides with linear and branched chain of carbon atoms containing 3 to 10 carbon atoms.

2. Preliminaries

Halogenated compounds represent a class of organic compounds with very diverse structures that depend on the nature of the halogen, the number of halogen atoms in the molecule, the nature of the hydrocarbon radical to which the halogen is bonded and the position of the halogen atoms in the chain. All these factors influence both the length and strength of the carbon-halogen chemical bond and the physical and chemical properties of these compounds. The nature of the hydrocarbon radical divides halogenated compounds into three different subclasses as to chemical behaviour and uses: alkyl halides, in which the halogen is bonded to an sp^3 hybridised saturated carbon atom, vinyl halides, in which the halogen is bonded to an sp^2 hybridised carbon atom of a double bond and aryl halides, in which the halogen bonds to a carbon atom of an aromatic nucleus.

Halogenated compounds are used in many different fields of activity: medicine, cosmetics, agriculture, dye industry, refrigeration industry, polymer industry. For example, vinyl chloride is a monomer of polyvinyl chloride, a polymer with very good mechanical and chemical resistance being one of the most used plastics, 2-chlorobutadiene is the monomer used to obtain artificial rubber, tetrafluoroethene is the monomer of Teflon, etc. Halogenated compounds are also of major importance for organic synthesis, being very important reaction intermediates. Some of the small-molecule halogenated compounds are used as solvents in synthesis, analysis laboratories and in industry: chloroform, methylene chloride, trichloroethene, etc. There are halogenated compounds used as alkylating agents, with applications in cancer chemotherapy.

Halogenated compounds are generally toxic. The use of halogenated compounds, especially pesticides (insecticides, herbicides, fungicides), raises environmental pollution problems as a result of their high stability to hydrolysis and resistance to biological, photolytic and chemical degradation. Due to their low solubility in water and high solubility in lipids, halogenated compounds accumulate in living organisms lipid stores, causing chronic ecotoxicity Gajewicz et al. [25], Xu et al. [50]. Freons, small chlorinated and chlorinated compounds of methane, used as propellant fluids in the form of aerosols, have negative effects by destroying the ozone layer Gajewicz et al. [25].

The physical, chemical and biological properties of halogenated compounds are influenced by the nature of the $C - X$ bond. The carbon-halogen bond is formed by the interpenetration of a p orbital of the halogen atom with a bonding orbital of a carbon atom (sp^3 in saturated compounds or sp^2 in unsaturated or aromatic ones). The volume of the orbitals of halogen atoms increases with increasing atomic volume of the element in the order $F < Cl < Br < I$. Due to the small volume of the carbon hybrid orbitals, the degree of interpenetration with halogen p orbitals decreases as their volume increases. Consequently, the $C - X$ interatomic distances increase and the bond energies decrease as the halogen order number increases. In unsaturated or aromatic halogenated compounds, the non-participating electrons of the halogen atoms interact with neighbouring π electrons giving rise to structures in which the carbon-halogen bond is partially double and shorter than in aliphatic halogens due to both $p - \pi$ conjugation and sp^2 hybridization of the atom of carbon.

If we consider the chemical structures C_i: C_1, C_2, C_3,..., C_n and the observed values of a certain property P denoted by y_i: $y_1, y_2, y_3, ..., y_n$, in QSPR / QSAR studies, we are interested in the development of a relationship between these properties and a descriptor of the respective structures. The linear regression equation is written as follows:

$$\hat{y}_i = y_{i,calc} = a + \sum_{j=1}^{m} b_j x_{ij},\ i = 1, 2, ..., n, \tag{2.1}$$

where

- a is a constant that is generally not assigned a physico-chemical meaning;
- b_j is the slope of the regression line, which is called the regression coefficient, $j = 1, 2, ..., m$;
- x_{ij} are independent variables also called predictor variables or explanatory variables, i.e., structural descriptors or topological indices that encode the structure of the compound C_i and condition the value of the measured property y_i;
- y_i are the observed values of the property P, also called the *dependent variable*.

The number m of independent variables indicates the type of regression, i.e., simple regression ($n = 1$) or multiple regression ($n \geq 2$).

The value of the constant a and the values of the regression coefficients b_j from the regression equation (2.1) are calculated by using the least squares method Beslau et al. [17]. According to this method, the condition is imposed so that between the observed values of the physico-chemical property or the biological activity, y and the values y calculated with equation (2.1), should be at a maximum compliance, a fact which is expressed by the property:

$$SS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} \left[y_i - \left(a + \sum_{j=1}^{m} b_j x_{ij} \right) \right]^2 = \text{minimum}, \tag{2.2}$$

where SS represents the sum of squares of the deviations of the values y_calc to the observed valuesy_i.

The difference between the observed property values and the calculated ones is called *residual* or *error*:

$$rest_i := y_{i,obs} - y_{i,calc}.$$

The function represented by equation (2.2) has a minimum when the partial derivatives with respect to a and b_j, respectively, are equal to zero. By equating with zero the partial derivatives with respect to $a, b_1, b_2, ..., b_n$ one obtains a system of equations from which the parameters $a, b_1, b_2, ..., b_n$ are determined.

The use of the regression equation as a QSPR/QSAR model is primarily conditioned by the quality of the statistical indicators: correlation coefficient, determination coefficient, standard deviation, Fisher F statistics, t-Student statistics etc., while the predictive value of the model is tested by applying different methods and techniques of internal validation and external validation.

The aim of this study is to develop statistical models with high goodness-of-fit, robustness and predictive ability for the boiling point and molar refraction of alkyl halides with Cl, Br and I. In order to as much as possible structural properties related to the degree of branching of chemical bonds, the hybridization state of the carbon atoms and the halides specificity, we used the topologic index ZEP as a main structural descriptor. ZEP is calculated by means of the notion of *weighted electronic distance*, introduced in [7] and have been successfully used as an independent variable to derive QSPR models for various classes of chemical compounds, see [7]-[16].

3. Molecular graphs and topological indices

3.1. **Modelling the structure of alkyl halides using molecular graphs.**

In general, a chemical compound is represented by a connected, non-oriented graph $G = (V, E)$, where $V = \{v_1, v_2, v_3, \ldots, x_n\}$ is the set of vertices or nodes (atoms), and $E = \{e_1, e_2, e_3, \ldots, e_m\}$ is the set of edges or arcs (chemical bonds between atoms).

The set E consists of non-oriented pairs $e = (v_i, v_j)$, called edges. Between any two atoms (nodes) there is at least one chain. Graphs that represent chemical structures are called chemical or *molecular graphs*. A basic characteristic of molecular graphs is that the degree of a node is at most equal to 4. In particular, the molecular graph of chemical compounds containing hetero-atoms is a marked graph.

Figure 1 shows five molecular graphs with five vertices and five edges representing the following compounds: 2-methylbutane (graph G(a), G(b)), 2-bromobutane (G1), 1-iodo-2-methylpropane (G2) and 2-chlorobutane (G3).

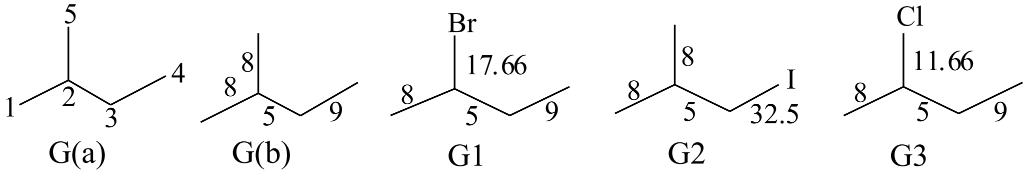


Figure 1. Molecualar graphs

The numbering and order of the vertices are entered, together with the topological distance between the vertices, in graph $G(a)$ from Figure 1, but these are the same for graphs $G(b)$, G_1, G_2 and G_3. In the graphs marked $G(b), G_1 - G_3$, we labelled edges with the weighted electronic distance, $wed(i, j)$, a concept introduced in Berinde [7].

The *weighted electronic distance*, $wed\,(i, j)$, is defined for each edge between two adjacent vertices in the graph, as:

$$wed\,(i, j) = \frac{1}{b_{ij}} \cdot \frac{Z_i' + Z_j'}{v_i \cdot v_j}, \tag{3.3}$$

where

- v_i is the degree of vertex i,
- $Z_i' = Z_i \cdot v_i$,
- Z_i denotes the order number of atom i,
- b_{ij} takes the values $1, 2, 3$ and 1.5 for a single bond, a double bond, a triple bond and an aromatic bond, respectively.

Z_i' represents a local vertex invariant (LOVI) in the molecular graph, while $w.e.d.(i, j)$ represents a local edge invariant (LOEI). The edge of graph is weighted with wed (weighted electronic distance).

For the most common bonds in alkyl halides, we calculated the values of $w.e.d.(i, j)$, which are given in Table 1.

TABLE 1. Values of wed for common single bonds in alkyl halides

Bond types	wed	Bond types	wed	Bond types	wed
$H_3C - CH_2$	9	$CH_2 - F$	10.5	$CH_2 - I$	32.5
$H_3C - CH$	8	$CH - F$	9	CH-I	23.66
$H_3C - C$	7.5	C-F	8.25	C-I	19.25
$CH_2C - CH_2$	6	$CH_2 - Cl$	14.5	$H_2C = CH$	2.5
$CH_2 - CH$	5	$CH - Cl$	11.66	$H_2C = C$	2.25
$CH_2 - C$	4.5	$C - Cl$	10.25	$HC = CH$	2
$CH - CH$	4	$CH_2 - Br$	23.5	$HC = C$	1.75
$CH - C$	3.5	$CH - Br$	17.66	$C = C$	1.5
$C - C$	3	$C - Br$	14.75	$C = C$	1

So, the weighted electronic distances for the edges in the graph $G(b)$ in Figure 1 are: $wed(1, 2) = 8$, $wed(2, 3) = 5$, $wed(3, 4) = 9$, and $wed(2, 5) = 8$.

The weighted electronic distance was used for the development of topological indices, which in turn are used for constructing QSPR / QSAR models of physico-chemical properties and biological activities.

3.2. Matrix representation of molecular graphs.

The adjacency matrix is the same for all compounds presented in Figure 1, i.e.,

$$A(G(a)) = A(G(b)) = A(G_1) = A(G_2) = A(G_3) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

By replacing the usual topological distances in the adjacency matrix by the weighted electron distances, we obtain the so called *weighted electron connectivity matrix*, denoted by CEP, which is an important source of topological indices, see Berinde [7].

In contrast to the case of adjacency matrix, which is the same for all structures in Figure 1, the CEP matrices corresponding to the same graphs are different from each other:

$$CEP(G(b)) = \begin{bmatrix} 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 5 & 0 & 8 \\ 0 & 5 & 0 & 9 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \end{bmatrix} ; CEP(G1) = \begin{bmatrix} 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 5 & 0 & 17.66 \\ 0 & 5 & 0 & 9 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 17.66 & 0 & 0 & 0 \end{bmatrix} ;$$

$$CEP(G2) = \begin{bmatrix} 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 5 & 0 & 8 \\ 0 & 5 & 0 & 32.5 & 0 \\ 0 & 0 & 32.5 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \end{bmatrix} ; CEP(G3) = \begin{bmatrix} 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 5 & 0 & 11.6 \\ 0 & 5 & 0 & 9 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 11.6 & 0 & 0 & 0 \end{bmatrix}$$

The sum of all entries on the i^{th} row in the CEP matrix is denoted by SEP_i:

$$SEP_i = \sum_{j=1}^{n} [CEP]_{ij},\ i = 1, 2, ..., n. \tag{3.4}$$

3.3. **The definition of the molecular descriptors. The structural parameter H_d**

To quantify the distances between the heteroatom and the other atoms in the graph, we introduced the structural parameter H_d, see Berinde [10], defined by the formula:

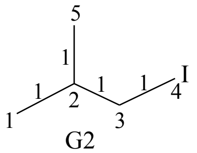
$$H_d = \frac{1}{n} \sum_{i=1}^{n} d_i, \tag{3.5}$$

where

- n represents the number of carbon atoms in the molecular graph;
- d_i represents the topological distance between the heteroatom and the atom i.

The structural parameter H_d can be calculated directly from the hydrogen-suppressed molecular graph or from the adiacency matrix.

The calculation technique of structural parameter H_d is illustrated in Figure 2 for the hydrogen-suppressed graph 1-bromo-2-methylpropane (G2).



$$H_d = \frac{1}{4}(3 + 3 + 2 + 1) = 2.25$$

FIGURE 2. Calculation of H_d for graph G2

We used the structural parameter H_d together the topological index ZEP as independent variables in multiple linear regression to develop quantitative structure-property model for estimating the normal boiling points and molar refractivity of alkyl halides.

The molecular topological index ZEP

The molecular topological index ZEP introduced by Berinde [7] is defined as:

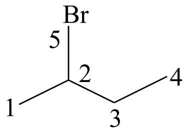
$$ZEP = \sum_{i=1}^{n} SEP_i^{\frac{1}{2}} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} wed(i,j) \right)^{\frac{1}{2}}, \tag{3.6}$$

where $wed(i,j)$ is the weighted electronic distance defined by (3.3).

3.4. **Steps in the computation of the ZEP index for a molecular graph.**

Below we indicate the steps in the computation of the ZEP index for a molecular graph representing the skeleton of 2-bromobutane (G1).

1. The labelling vertices in the hydrogen-suppressed molecular graph:

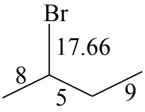


2. Calculation of weighted electronic distances for edges:

$$wed_{12} = \frac{1 \times 6 + 3 \times 6}{1 \times 3} = 8, \quad wed_{23} = \frac{3 \times 6 + 2 \times 6}{3 \times 2} = 5,$$

$$wed_{34} = \frac{2 \times 6 + 1 \times 6}{2 \times 1} = 9, \quad wed_{25} = \frac{3 \times 6 + 1 \times 35}{3 \times 1} = 17.66.$$

3. Labelling the edges in the hydrogen-suppressed molecular graph:



4. The adjacency matrix and CEP matrix of graph (G1):

$$A(G1) = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}; CEP(G1) = \begin{bmatrix} 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 5 & 0 & 17.66 \\ 0 & 5 & 0 & 9 & 0 \\ 0 & 0 & 9 & 0 & 0 \\ 0 & 17.66 & 0 & 0 & 0 \end{bmatrix}$$

5. The computation of ZEP index of graph G1

$$ZEP(G1) = 8^{1/2} + 30.66^{1/2} + 14^{1/2} + 9^{1/2} + 17.66^{1/2} = 20.616.$$

TABLE 2. The training set: the H_d parameter, the ZEP indices of alkyl halides and their boiling points (*bp*) and molar refraction (mr)

No.	compound	N	Hd	ZEP	*bp* (° C)	*bp* (° C)	mr	mr
					$exp.$	$calc.$	$exp.$	$calc$
1	1-Cl-propane	3	2.000	15.208	48.2	56.08	20.56	20.63
2	1-Br-propane	3	2.000	17.152	74.3	83.3	23.98	24.71
3	1-I-propane	3	2.000	18.778	104.5	106.1	28.64	28.13
4	2-Cl-propane	3	1.667	14.322	38.7	43.6	20.95	20.57
5	2-Br-propane	3	1.667	15.660	62.5	62.4	23.36	23.37
6	2-I-propane	3	1.667	16.818	89.5	78.6	28.61	25.81
7	1-Cl-butane	4	2.500	18.672	79.2	80.4	25.67	25.22
8	1-Br-butane	4	2.500	20.616	103.6	107.6	28.72	29.29
9	1-I-butane	4	2.500	22.242	130.5	130.4	33.61	32.71
10	2-Cl-butane	4	2.000	17.950	70.8	70.3	25.63	26.38
11	2-Br-butane	4	2.000	19.309	91.2	89.3	28.64	29.23
12	2-I-butane	4	2.000	20.488	118.0	105.8	33.51	31.72
13	1-Cl-2-methylpropane	4	2.250	18.463	69.5	77.5	25.61	26.12
14	1-Br-2-methylpropane	4	2.250	20.425	92.2	104.9	28.67	30.23
15	1-I-2-methylpropane	4	2.250	22.064	123.4	127.9	33.725	33.67
16	2-Cl-2-methylpropane	4	1.750	17.140	53.0	58.9	25.46	26.03
17	2-Br-2-methylpropane	4	1.750	18.159	73.3	73.2	28.18	28.17
18	2-I-2-methylpropane	4	1.750	19.065	97.0	85.9	33.17	30.07
19	1-Cl-pentane	5	3.000	22.037	106.4	104.7	30.54	29.80
20	1-Br-pentane	5	3.000	24.080	132.0	131.9	33.27	33.87
21	1-I-pentane	5	3.000	25.706	156.0	154.7	38.06	37.28
22	2-Cl-pentane	5	2.400	21.398	94.3	94.3	30.32	31.47
23	2-Br-pentane	5	2.400	22.757	114.3	113.4	33.12	34.32
24	2-I-opentane	5	2.400	23.936	138.2	129.9	37.87	36.79
25	3-Cl-pentane	5	2.200	21.552	97.3	96.5	30.43	32.86
26	3-Br-pentane	5	2.200	22.944	116.5	116.0	33.18	35.78
27	3-I-pentane	5	2.200	24.149	141.4	132.9	37.93	38.31
28	1-Cl-hexane	6	3.500	25.601	130.2	129.0	34.47	34.37
29	1-Br-hexane	6	3.500	27.544	156.4	156.2	37.92	38.45
30	1-I-hexane	6	3.500	29.171	179.5	179.0	42.58	41.86
31	1-Cl-heptane	7	4.000	29.065	158.3	153.3	39.01	38.95
32	1-Br-heptane	7	4.000	31.008	184.6	180.6	42.64	43.02
33	1-I-heptane	7	4.000	32.635	206.0	203.4	47.05	46.44
34	1-CL-octane	8	4.500	32.529	179.0	177.6	43.71	43.53
35	1-Br-octane	8	4.500	34.473	207.8	204.9	47.13	47.60
36	1-I-heptane	8	4.500	36.099	227.5	227.7	51.88	51.01
37	1-Cl-octane	9	5.000	35.993	203.5	202.0	48.30	48.11
38	1-Br-nonane	9	5.000	37.936	226.6	229.2	51.82	52.18
39	1-I-nonane	9	5.000	39.563	248.3	252.0	56.16	55.59
40	1-Cl-decane	10	5.500	39.457	226.0	226.3	52.91	52.68
41	1-Br-decane	10	5.500	41.400	251.5	253.5	56.44	56.76
42	1-I-nodecane	10	5.500	43.027	268.2	276.3	59.98	60.17

4. QSPR MODELS FOR NORMAL BOILING POINT AND MOLAR REFRACTION

4.1. The dependent and independent variables used in the study.

The physicochemical properties of alkyl halides selected in this study were: molar refraction (or molar refractivity), denoted by mr and boiling point, denoted by bp. Experimental data sets for these properties are obtained from literature: Kier and Hall [30], Lide [32], [51], and are listed in Tables 2 and 3, along with the name of the compounds.

The data set comprises 64 alkyl halides including clorohalides, bromohalides and iodohalides.

The data set was randomly divided into two subsets, called the training set, with 42 compounds used to obtain the QSPR equation (Table 2) and the validation set with 22 compounds used for external validation (Table 3).

Thus we calculated for each of the 64 halogenated compounds, the topological index ZEP and the structural parameter H_d. These values can be found in Tables 2 and 3. ZEP index and the structural parameter H_d were then used as independent variables in simple and multiple linear regression to construct structure-property QSPR models.

The SPSS version 20 software package was used to obtain the regression equation.

4.2. **QSPR equations.**

Using the simple and multiple linear regression method as a statistical tool for developing reliable QSPR models, the following linear model was obtained, in which the molecular descriptors ZEP, Hd and N (the number of carbon atoms), were used as independent variables:

Regression equation and statistical parameters for monovariable correlation of boiling points with the ZEP index are shown below:

$$bp = -55.607 + 7.575 \cdot ZEP \tag{4.7}$$

$$N = 42;\ R = 0.983;\ R^2 = 0.967;\ R^2_{adj} = 0.966;\ s = 11.2;\ F = 1175.069;\ sig = 0.000.$$

The statistical results of equation (4.7) show that ZEP index have good correlation with boiling points in terms of the coefficient of determination, the standard error and the Fisher statistic value. A plot of the boiling points values versus the values of ZEP index, shows a direct linear correlation (Figure 4).

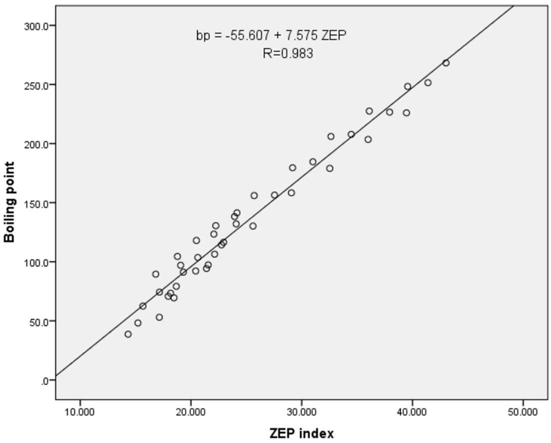


FIGURE 3. Plot of boiling points of alkyl halides versus topological index

Regression equation and statistical parameters for two-variable correlation of boiling points with the ZEP index and the structural parameter, H_d are shown below:

$$bp = -72.581 + 10.840 \cdot ZEP - 21.587 \cdot H_d \tag{4.8}$$

$$N = 42;\ R = 0.988;\ R^2 = 0.977;\ R^2_{adj} = 0.975;\ s = 9.6;\ F = 814.484;\ sig = 0.000.$$

Regression equation and statistical parameters for two-variable correlation of boiling points, with the ZEP index and the number of carbon atoms, N are shown below:

$$bp = -84.350 + 14.012 \cdot ZEP - 24.219 \cdot N \tag{4.9}$$

$$N = 42; \quad R = 0.996; \quad R^2 = 0.992; \quad R^2_{adj} = 0.992; \quad Q^2_{ext} = 0.991;$$
$$s = 5.4; \quad F = 2572.637; \quad R^2_{CV} = 0.9898; \quad sig = 0.000.$$

These QSPR models present high values correlation coefficient ($R > 0.98$) Fisher-ratio statistics. The best statistical results show the model (4.9), where it was used the combination of ZEP index and the number of carbon atoms, N. The boiling points calculated by equation (4.9) are listed in Table 3.

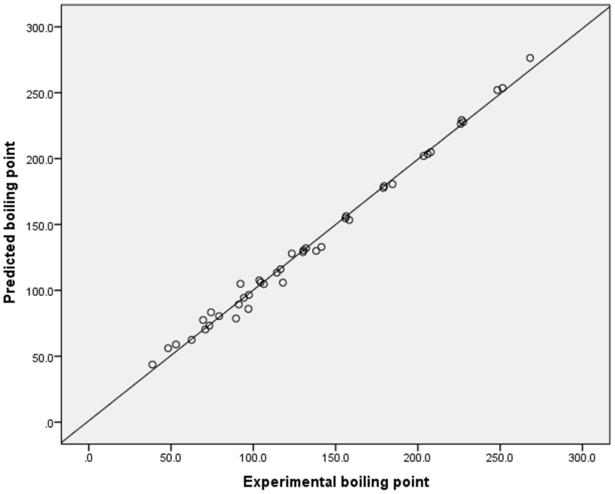


FIGURE 4. Plot of predicted versus experimental values of alkyl halides boiling points

Regression equation and statistical parameters for monovariable correlation of molar refraction (mr) with the ZEP index are shown below:

$$mr = 3.715 + 1.285 \cdot ZEP \tag{4.10}$$

$$N = 42; \; R = 0.984; \; R^2 = 0.969; \; R^2_{adj} = 0.968; \; s = 1.8; \; F = 1229.977; \; sig = 0.000$$

The statistical results of equation (4.10) shows that the ZEP index have good correlation with molar refraction in terms of the coefficient of determination, the standard error and the Fisher statistic value. A plot of the molar refraction values versus ZEP index, shows a direct linear correlation (Figure 5).

Regression equation and statistical parameters for two-variable correlation of molar refraction with the ZEP index and the number of carbon atoms, N, are shown below:

$$mr = 0.946 + 1.905 \cdot ZEP - 2.333 \cdot N \tag{4.11}$$

$$N = 42; \; R = 0.988; \; R^2 = 0.977; \; R^2_{adj} = 0.976; \; s = 1.62; \; F = 817.893; \; sig = 0.000$$

The regression equation and statistical parameters for two-variable correlation of molar refraction with the ZEP index and the structural parameter, H_d, are shown below:

$$mr = -0.510 + 2.097 \cdot ZEP - 5.374 \cdot H_d \tag{4.12}$$

$$N = 42; \quad R = 0.995; \quad R^2 = 0.989; \quad R^2_{adj} = 0.989; \quad Q^2_{ext} = 0.990; \; s = 1.1;$$
$$F = 1768.212; \; R^2_{CV} = 0.9875; \; sig = 0.000$$

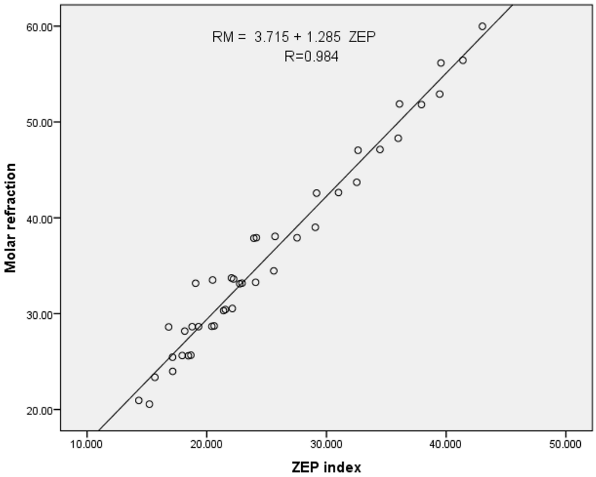


FIGURE 5. Plot of molar refraction versus ZEP index for alkyl halides

These QSPR models present high values correlation coefficient ($R > 0.98$) Fisher-ratio statistics. The best statistical results show the model 9, where it was used the combination of ZEP index and the structural parameter, H_d. The molar refraction values calculated by equation (4.12) are listed in Table 3.

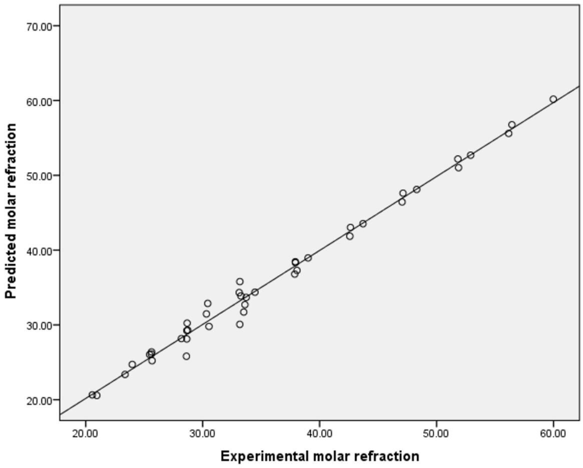


FIGURE 6. Plot of experimental versus predicted values of molar refraction

TABLE 3. The external validation set - the H_d parameter, the ZEP index of alkyl halides and their boiling points (bp) and molar refraction (mr)

No.	compound	H_d	N	ZEP	bp(° C) *obs.*	bp(° C) *calc.*	mr *obs.*	mr *calc.*
1	2-cloro-3-methylbutane	2.200	5	21.518	95.4	96.1	30.3	32.78
2	1-bromo-2-methylbutane	2.600	5	23.999	131.2	130.8	33.2	35.83
3	1-cloro-3-methylpentane	3.166	6	25.465	128.5	127.1	34.30	35.86
4	2-clorohexane	2.833	6	24.863	121.0	118.7	34.70	36.39
5	2-cloro-2-methylpentane	2.333	6	24.246	112.4	110.1	34.60	37.79
6	2-cloro-3-methylpentane	2.500	6	24.800	117.5	117.8	34.65	38.05
7	3-cloro-2-methylpentane	2.333	6	24.831	118.6	118.3	34.62	39.01
8	2-bromo-2-methylpentane	2.333	6	25.284	133.1	124.6	37.30	39.96
9	3-iodohexane	2.500	6	27.597	164.5	157.1	42.60	43.92
10	2-cloroheptane	3.286	7	28.327	145.0	143.0	39.10	41.22
11	2-bromo-2-methylhexan	2.714	7	28.748	156.6	148.9	42.20	45.18
12	3-iodoheptane	2.857	7	31.061	187.2	181.3	47.10	49.26
13	1-cloro-3-methylheptane	4.000	8	32.377	173.4	175.6	43.70	45.87
14	2-clorooctane	3.750	8	31.791	169.1	167.4	43.90	45.99
15	2-cloro-2-methylheptane	3.125	8	31.175	161.9	158.7	43.80	48.06
16	1-bromo-2-methylheptane	3.875	8	34.375	198.8	203.6	47.00	50.73
17	3-iodooctane	3.250	8	34.525	211.2	205.7	52.00	54.41
18	3-iodo-2-methylheptane	2.875	8	34.362	207.3	203.4	51.90	56.08
19	2-clorononane	4.222	9	35.255	195.5	191.7	48.30	50.71
20	3-iodononane	3.666	9	37.989	232.4	230.0	56.50	59.44
21	2-clorodecane	4.700	10	38.719	216.7	216.0	52.90	55.41
22	2-cloro-2-methyldecane	4.454	11	41.567	233.6	231.7	58.10	62.70

4.3. **Validation of the QSPR models.**

The robustness of the QSPR models (4.9) and (4.12) as well as their internal predictive ability were evaluated by R^2_{CV} - cross validation coefficient based on leave-one-out (LOO), external validation coefficient Q^2_{ext} and y-randomization test, see Topliss and Costello [45], Tropsha et al. [48].

Y-randomization test was performed 10 times. In each y-randomization run, $R^2_{yi} < 0.2$, which shows that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

TABLE 4. Ten y-randomizations for PLS model on data set

Randomization (boiling point)	$R^2_{y_i}$	Randomization (molar refraction)	$R^2_{y_i}$
1	0.111	1	0.011
2	0.103	2	0.008
3	0.005	3	0.024
4	0.097	4	0.012
5	0.034	5	0.083
6	0.039	6	0.014
7	0.095	7	0.033
8	0.032	8	0.108
9	0.076	9	0.019
10	0.053	10	0.060
maximum value	**0.111**	maximum value	**0.108**
average	**0.0635**	average	**0.0372**

5. Conclusions

In this study, various QSPR mathematical models have been developed to predict the normal boiling point and the molar refractivity of a larger set of alkyl halides by using multilinear regression methods based on the following topological descriptors: the ZEP topological index (based on the weighted electronic distances), the structural parameter H_d, and the number of carbon atoms (N) as independent variables.

These quantitative relationships are of great technological importance since, based on them, one can predict the properties of new untested molecules by means of QSPR functional expression obtained, and were obtained by performing specific QSPR studies on a class of similar compounds whose properties have been already determined and were appropriately correlated to their molecular structures.

We determined three QSPR models for the normal boiling point (*bp*), the best of them being given by (4.9), with $R = 0.996$ and $s = 5.4$.

For the study of molar refraction (mr) we developed three QSPR models, of which the best one is given by (4.12), with $R = 0.995$ and $s = 1.1$.

We have validated the developed models with respect to the goodness-of-fit, robustness and predictive ability for the two properties considered in the class of alkyl halides, external validation, cross-validation (leave-one-out) and randomization (y-randomization) were used.

The obtained results have shown that the three descriptors (ZEP, H_d and N) could be efficiently used for modelling and predicting the the normal boiling points and molar refractions of the considered class of compounds.

For other related developments that could be approached by means of the tools used in the current paper, we refer to Baumann and Baumann [6], Consonni et al. [19], Ghosh et al. Ghosh, Kiralj and Ferreira [31], Liu and Long [33], Padrón et al. [37], Pogliani [38], Roy et al. [39], [40], Rücker et al. [42].

References

[1] Arjmand, F.; Shafiei. F. Prediction of the normal boiling points and enthalpy of vaporizations of alcohols and phenols using topological indices. *J. Structural Chem.* **59** (2018), no. 3, 748–754.

[2] Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between chemical structure and normal boiling points of halogenated alkanes C1-C4. *J. Chem. Inf. Comput. Sci.* **32** (1992), no. 3, 233–237.

[3] Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D. Correlation between Structure and Normal Boiling Points of Haloalkanes C1-C4 Using Neural Networks, *J. Chem. Inf. Comput. Sci.* **34** (1994), 1118–1121.

[4] Barysz, M.; Jashari, G.; Lall, R. S.; Srivastaya, V. K.; Trinajstić, N. On the distance matrix of molecules containing heteroatoms, in *Chemical Applications of Graph Theory and Topology*, King, R. B. (Ed), Elsevier, Amsterdam, 1983, 222–230.

[5] Basak, S. C.; Gute, B. D.: Grunwald, G. D. Estimation of the Normal Boiling Points of Haloalkanes Using Molecular Similarity, *Croat. Chem. Acta* **69** (1996), 1159–1173.

[6] Baumann, D.; Baumann, K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation, *J. Cheminf.* **6** (2014), Article no. 47, 1–19.

[7] Berinde, Z. *Applications of Molecular Topology in the Study of Physico-chemical Properties of Organic Compounds* (in Romanian), Cub Press 22, Baia Mare, 2001.

[8] Berinde, Z. Consideraţii privind modelarea matricială a compuşilor halogenaţi, *Rev. Chim. (Bucureşti)*, **52** (2001), no. 12, 788–792.

[9] Berinde, Z. QSPR modelling of molar volume of alkanes using the ZEP topological index, *Creat. Math. Inform.* **17** (2008), no. 3, 308–312.

[10] Berinde, Z., Using the topological index ZEP in QSPR studies of alcohols, *Studia Univ. Babeş-Bolyai, Chemia*, 54 (2009), no. 4, 152–163.

[11] Berinde, Z., Matrix mathematical models used in the representation of molecular structures, *Sc. Stud. Res. Ser. Math. Inf.* **19** (2009), no. 2, 59–70.

[12] Berinde, Z., Modelling normal boiling points of alkanes by linear regression using the SD index, *Creat. Math. Inform.* **19** (2010), no. 2, 135–139.

[13] Berinde, Z., A QSPR study of hydrophobicity of phenols and 2-(aryloxy-α-acetyl)-phenoxathiin derivatives using the topological index ZEP, *Creat. Math. Inform.* **22** (2013), no. 1, 33–40.

[14] Berinde, Z., QSTR mathematical models for the toxicity of aliphatic carboxylic acids on tetrahymena pyriformis, *Creat. Math. Inform.* **22** (2013), no. 2, 151–160.

[15] Berinde, Z., Comparing the molecular graph degeneracy of Wiener, Harary, Balaban, Randić and ZEP topological indices, *Creat. Math. Inform.* **23** (2014), no. 2, 165–174.

[16] Berinde, Z.; Butean, C.; Dippong, T. Development of a QSPR model for predicting octane number of alkanes using SD topological index. *Creat. Math. Inform.* **25** (2016), no. 2, 151–158.

[17] Besalu, E.; de Julian-Ortiz, J. V.; Pogliani, L. Trends and plot methods in MLR studies, *J. Chem. Inf. Model.* **47** (2007) 751–760.

[18] Carlton, T. S. Correlation of Boiling Points with Molecular Structure for Chlorofluoroethanes, *J. Chem. Inf. Comput. Sci.* **38** (1998), 158–164.

[19] Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* 24 (2010), 194–201.

[20] Dai Y.-M.; Zhu Z.-P.; Cao Z.; Zhang Y.-F., Zeng J.-L., Li X. Prediction of boiling points of organic compounds by QSPR tools. J. Molecular Graphics Model. 44 (2013), 113–119.

[21] Diudea, M. V.; Ivanciuc, O. *Molecular Topology* (in Romanian), Comprex, Cluj-Napoca, 1995.

[22] Diudea, M. V.; Gutman, I.; Jantschi, L. *Molecular Topology*, Nova Science Pub Inc; UK ed. edition, 2001.

[23] Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; Gonzalez, M. P. A new search algorithm for QSPR/QSAR theories: Normal boiling points of some organic molecules, *Chem. Phys. Lett.* **412** (2005), no. 4-6, 376–380.

[24] Engel, T.; Gasteiger, E. J. *Applied Chemoinformatics: Achievements and Future Opportunities*, John Wiley & Sons, Weinheim, Germany, 2018.

[25] Gajewicz, A.; Haranczyk, M.; Puzyn, T. Predicting logarithmic values of the subcooled liquid vapor pressure of halogenated persistent organic pollutants with QSPR: How different are chlorinated and brominated congeners? *Atmospheric Environment* **44** (2010), 1428–1436.

[26] Garcia-Domenech, R.; Galvez, J.; de Julian-Ortiz, J. V.; Pogliani, L. Some new trends in chemical graph theory, *Chem. Rev.* **108** (2008), 1127–1169.

[27] Gharagheizi F.; Mirkhani, S. A.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D.; Richon, D. Determination of the normal boiling point of chemical compounds using a quantitative structure-property relationship strategy: Application to a very large dataset. *Fluid Phase Equil.* **354** (2013), 250–258.

[28] Ghosh, S.; Ojha, P. K.; Roy, K. Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs, *Chemosphere* **228** (2019), 545–555.

[29] Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Quantitative Structure-Property Relationship Study of Normal Boiling Points for Halogen-/ Oxygen-/ Sulfur-Containing Organic Compounds Using the CODESSA Program, *Tetrahedron* **54** (1998), 9192–9142.

[30] Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.

[31] Kiralj, R.; Ferreira, M. M. C. Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application. *J. Braz. Chem. Soc.* 20 (2009), 770–787.

[32] Lide, D. R. *CRC Handbook of Chemistry and Physics*, 83rd ed., CRC Press, Boca Raton, FL, 2002–2003.

[33] Liu P. X.; Long, W. Current Mathematical Methods Used in QSAR/QSPR Studies, *Int. J. Mol. Sci.* **10** (2009), 1978–1998.

[34] Lu, H.; Yang, F.; Liu, W.; Yuan, H.; Jiao, Y. A robust model for estimating thermal conductivity of liquid alkyl halides, *SAR and QSAR Environmental Res.* **31** (2020), no. 2, 73–85.

[35] Mihalić, Z.; Trinajstić, N. A graph-theoretical approach to structure-property relationships, *J. Chem. Educ.* 69 (1992), no. 9, 701–712.

[36] Öberg, T. Boiling points of halogenated aliphatic compounds: a quantitativestructure-property relationship for prediction and validation, *J. Chem. Inf. Comput. Sci.* **44** (2004) 187–192.

[37] Padrón, J. A.; Carrasco, R.; Pellón, R.F. Molecular descriptor based on a molar refractivity partition using Randić-type graph-teoretical invariant, *J. Pharm. Pharmaceut. Sci.* **5** (2002), no. 3, 258–266.

[38] Pogliani, L. Model of the physical properties of halides with complete graph-based indices, *Int. J. Quant. Chem.* **102** (2005) 38–52.

[39] Roy, K.; Kar, S.; Das, R.N. *Statistical Methods in QSAR/QSPR*, Springer International Publishing, Cham, 2015.

[40] Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H. Comparative studies on some metrics for external validation of QSPR models, *J. Chem. Inf. Model.* **52** (2012), 396–408.

[41] Roy, P. P.; Paul, S.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive QSAR models, *Molecules* **14** (2009), no. 4, 1660–1701.

[42] Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **47** (2007), 2345–2357.

[43] Sanghvi, R.; Yalkowsky, S. H. Estimation of the normal boiling point of organic compounds. *Ind. Eng. Chem. Res.* **45** (2006) 2856–2861.

[44] Sola, D.; Ferri, A.; Banchero, M.; Manna, L.; Sicardi S. QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method. *Fluid Phase Equil.* **263** (2008), no. 1, 33–42.

[45] Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* 15 (1972), no. 10, 1066–1068.

[46] Trinajstić, N. *Chemical Graph Theory.* CRC Press, Inc. Boca Raton, Florida, 1983

[47] Trinajstić, N. *Chemical Graph Theory.* CRC Press, Boca Raton, 1992.

[48] Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, *QSAR Comb. Sci.* **22** (2003), 69–77.

[49] Wei, J. Boiling Points and Melting Points of Chlorofluorocarbons, *Ind. Eng. Chem. Res.* **39** (2000), 3116–3119.

[50] Xu, H.-Y.; Zhang, J.-Y.; Zou, J.-W.; Chen, X.-S. QSPR models for the physicochemical properties of halogenated methyl-phenyl ethers, *J. Molecular Graph. Model.* **26** (2008) 1076–1081.

[51] http://www.chemspider.com

[1]TECHNICAL UNIVERSITY OF CLUJ-NAPOCA
FACULTY OF SCIENCES NORTH UNIVERSITY CENTER AT BAIA MARE
DEPARTMENT OF CHEMISTRY AND BIOLOGY
76 VICTORIEI STREET, 430122 BAIA MARE, ROMANIA
Email address: zoita.berinde@cb.utcluj.ro
Email address: claudia.butean@cb.utcluj.ro