Polish J. Chem., **77**, 1491–1499 (2003)

Prediction of Boiling Points of Ketones Using a Quantitative Structure-Property Relationship Treatment*

by A. Komasa

Department of Chemistry, A. Mickiewicz University,
Grunwaldzka 6, 60-780 Poznań, Poland; e-mail: aniak@amu.edu.pl

(Received April 1st, 2003; revised manuscript June 3rd, 2003)

A data set of 202 ketones was used to study a correlation of over 700 descriptors with boiling point using a Quantitative Structure-Property Relationship (QSPR) technique. The QSPR study was performed with the CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) program. A large part of the descriptors used in the study, was based on quantum-chemical calculations. A four parameter model with the squared correlation coefficient $R^2 = 0.9907$ and the Fisher ratio F = 5231 was obtained. The final QSPR model includes one topological and three quantum-chemical descriptors: Kier&Hall index (order 1), Total molecular two-center resonance energy divided by the number of atoms, Maximum net atomic charge for a C atom (Gaussian NBO), and Average valency of a C atom. The four-descriptor relationship can be recommended for prediction of the boiling point of ketones.

Key words: Quantitative Structure-Property Relationship, molecular descriptors, property prediction, ketones

Knowledge of the physical properties of organic compounds is necessary for the design, development and manufacture of products in which they are used. The suitability of a particular compound for a given purpose depends on its physicochemical properties. The boiling point is one of the main physicochemical properties used to characterize and identify compounds. It is predetermined by the intermolecular interactions in the liquid and by the difference in the molecular internal partition function in the gas and liquid phase at the boiling point. So, it depends indirectly on the chemical structure of the interacting molecules and it is not surprising that numerous methods have been developed for estimating the boiling point of a compound from its structure [1]. One of the most widespread is the Quantitative Structure-Property Relationship (QSPR) technique, which has been successfully applied in physical, organic, analytical, pharmaceutical and medicinal chemistry, biochemistry, chemical engineering and technology, toxicology and environmental sciences for the past twenty years [2,3]. The wide application of the QSPR is based on the possibility of estimating the properties of new chemical compounds without the need to synthesize and test them. The assumption of the

* Dedicated to Prof. M. Szafran on the occasion of his 70th birthday.

QSPR method is that a given property (physical, chemical and biological) correlates with the molecular structure of an individual compound. In a numerical form the chemical structure is represented by various theoretical descriptors, which reflect simple molecular properties and thus can provide insight into the physicochemical nature of the property or activity under consideration. Pioneering work in applying the QSPR to boiling point was done by Wiener, who introduced the path number *w* (later named the Wiener index), defined as the sum of the distances between any two carbon atoms in the molecule [4]. Since then, extensive effort has been made to apply the structural information to fit experimental boiling points. In most instances, correlations have been made initially for organic compounds in homologous series, like: furans, tetrahydrofurans and thiophenes [5], pyrans and pyrroles [6], heterocyclic compounds [7,8,9], fluorocarbons [10], aliphatic hydrocarbons [11], and later have been extended to a diverse set of compounds [12,13,14].

In this study the QSPR technique was employed to estimate the normal boiling point (the boiling point at 760 mm Hg) of ketones. The QSPR analysis was performed using the CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis) program [15]. CODESSA is a chemical multi-purpose statistical analysis program that has already been successfully applied by Katritzky and co-workers to correlate molecular structure with different properties such as boiling point, melting point, critical temperature, vapour pressure, refractive index, critical micelle concentrations and interactions between different molecular species (octanol-water partition coefficient, aqueous solubility of liquids, solid and gases, solvent polarity scales, GC retention time and response factor) [16]. Although numerous attempts have been made to correlate different physical properties of organic compounds (among them boiling point) with structural parameters, there are very few papers employing quantum-chemical descriptors for this purpose. In this study a large part of descriptors is based on quantum-chemical calculations including purely *ab initio* descriptors. These kinds of descriptors allow a far more precise and complete insight into electronic structure of molecules and into intermolecular interactions. Moreover, extension of the collection of descriptors, in the present study, by adding almost a hundred and eighty quantum-chemical descriptors, increases the possibility to find descriptors well correlating with boiling points.

METHODOLOGY

QSPR study was derived following the methodology shown in Fig. 1. First, 202 ketones and the corresponding experimental boiling point data were selected from available compilations in literature (Beilstein, Aldrich, [17]). The set of ketones included aliphatic, aromatic and cyclic ketones, whose exemplary structures are given in Fig. 2. The boiling point data covered a wide range of temperatures from 79 to 420°C. For each ketone the structural formula was converted to a 3D representation using a standard set of bond lengths and angles. Such a preliminary geometry was optimized in quantum-chemical calculations at a semiempirical level of theory.

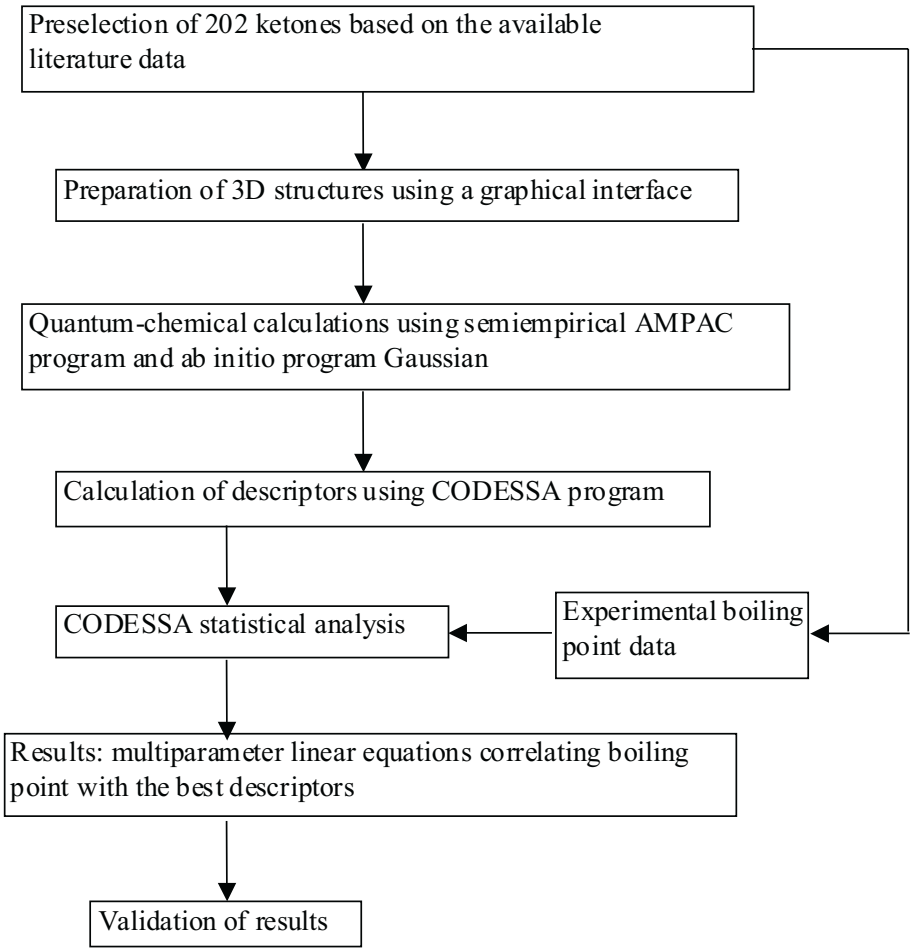

Figure 1. Process flow diagram for developing QSPR for predicting boiling points of ketones.

The full geometry optimization was performed by means of AMPAC Ver. 6.55 program employing AM1 Hamiltonian [18]. Such an optimized geometry was subsequently used as input data for three separate quantum-chemical calculations. In the first run the population analysis was computed at the AM1 level in order to the supply data on the electron charge density distribution, orbital populations, bond orders and valences, dipole moments, dipole and higher polarizabilities and on the partitioning of the energy into one- and two-center contributions. In the second run the thermodynamic properties were obtained at the same level of theory. From this run the properties like enthalpy, entropy, heat capacity, and zero point vibrational energy were derived. Finally, the third run of calculations was performed at *ab initio*

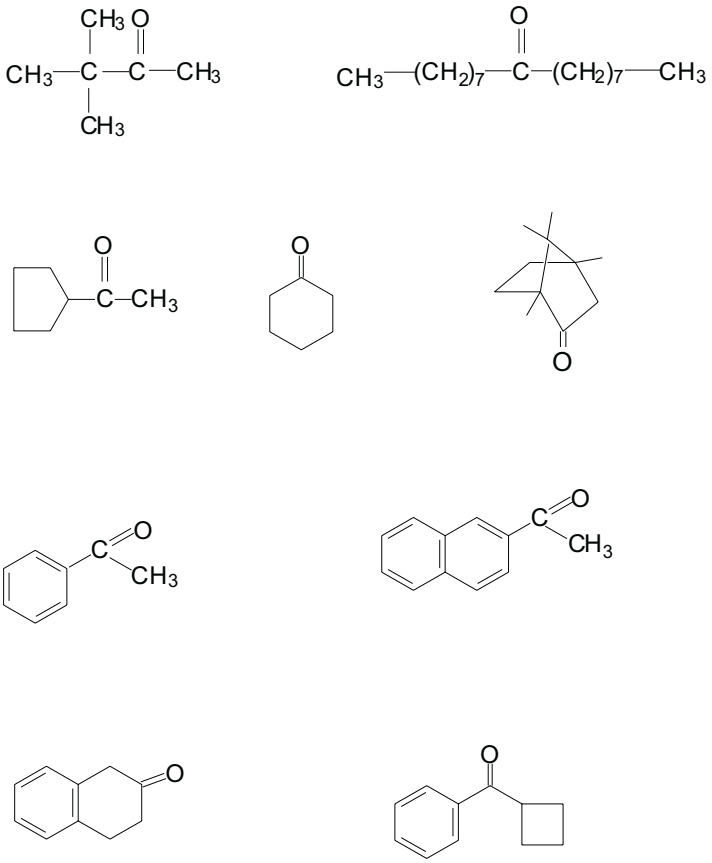

Figure 2. Exemplary structures of ketones used to develop the boiling point model.

level using Hartree-Fock method and 6-31G basis set, as implemented in Gaussian 98 package [19]. A large variety of properties based on the electron density distribution were extracted from this run. In the next step the outputs from these three runs of computations served as the input data for the CODESSA program, which carried out the calculations of molecular descriptors. The molecular descriptors generated by the CODESSA program can be divided into six groups: constitutional, topological, geometrical, electrostatic, quantum-chemical and thermodynamic. Constitutional descriptors are based on the molecular composition of the compound. These descriptors are simply counts of atoms, bonds, rings, molecular weight, and so on. Topological descriptors characterize the atomic connectivity in the molecule [20] and they are represented by well-known molecular parameters like the Wiener index [4], the Randic indices [21], Kier&Hall connectivity indices [22], Kier shape flexibility indices [23], *etc.* Geometrical descriptors represent the three-dimensional characteristics of the molecular structure such as *e.g.* moments of inertia [24],

shadow indices [25], molecular volume and molecular surface area [26,27]. The electrostatic descriptors characterize the charge distribution in the molecule. Among them, there are the empirical partial charges in the molecule calculated using the approach proposed by Zefirov [28], maximum and minimum atomic partial charges, polarity parameter [29] and topological electronic index [30]. CODESSA also uses a set of 26 Charged Partial Surface Area (CPSA) descriptors introduced by Jurs *et al*. [31,32]. Thermodynamic descriptors are calculated quantum-chemically on the basis of the thermodynamic partition function of the molecule. Quantum-chemical descriptors are a relatively new group of molecular descriptors, adding important information to the conventional descriptors. The most frequently used quantum-chemical descriptors include the energy of the highest occupied and lowest unoccupied molecular orbitals, frontier orbital electron densities, Mulliken population charge distribution, dipole moments and polarizabilities. A more detailed description and the application of quantum-chemical descriptors in QSPR study have been reviewed in [33].

The initial data for statistical analysis were composed of the boiling points (BP) and over 700 descriptors for each out of 202 ketones. In the preliminary heuristic treatment the descriptors were screened for insignificance, missing values and high intercorrelation, which strongly reduced the number of descriptors for further study. Such a limited set of descriptors was employed in the search for the best fits of the multiparameter linear equation

$$\mathrm{BP} = a_0 + \sum_i a_i d_i \tag{1}$$

where a_i are parameters of the fit characterizing a relative sensitivity of the boiling point (BP) to a given descriptor, d_i. In this way the best three-descriptor correlation was found. This correlation was successively improved giving finally the best four- and five-descriptor equations.

RESULTS AND DISCUSSION

The final models were extracted on the basis of the highest values of squared correlation coefficient (R^2) and the highest Fisher test values (F). The squared correlation coefficient is a measure of the quality of the linear relationship and the F-test value represents the completeness of the fit. For each model the standard deviation (s) and the cross-validated correlation coefficient (R_{CV}), which is essentially a characteristic of the predictive power of the correlation equation, were also calculated.

Details of the improvement of the statistical description of the boiling point data set (according to the square of the regression correlation coefficient R^2 and the F-values) with the increasing number of descriptors involved in the correlation are given in Table 1.

Table 1. Improvement of the statistical description of the boiling point data set with increasing number of descriptors involved in the correlations.

n	R^2	F	s	R^2_{CV}	Descriptor, d_i
1	0.6746	414	37.4	0.6674	Kier&Hall index (order1)
2	0.9793	4704	9.5	0.9785	Total molecular two-center resonance energy/number of atoms
3	0.9859	4616	7.8	0.9849	Maximum net atomic charge for a C atom (Gaussian NBO)
4	0.9907	5231	6.4	0.9901	Average valency of a C atom
5	0.9924	5151	5.8	0.9917	FPSA-2 Fractional PPSA (PPSA-2/TMSA) [Gaussian NBO PC]

The four-descriptor model yields a correlation characterized by a very satisfactory value of $R^2 = 0.9907$. The standard deviation of the model is s = 6.4°C. The cross-validated correlation coefficient $R^2_{CV} = 0.9901$ does not practically differ from the correlation coefficient R^2 which indicates a stability of the obtained QSPR model. Extension of the model by another descriptor improves R^2 only slightly but simultaneously worsens the F-test value. This is an indication that the four-descriptor model carries out the most essential information on the system.

The descriptors employed in the final model are shortly described below in order to facilitate their full recovery. The full list of compounds employed in this study along with their experimental boiling points and the four descriptors involved in the final correlation is available from the author on request.

The first descriptor involved in the model is the topological *Kier&Hall (order 1)* index

$$^1\chi = \sum_{bonds} (\delta_i \delta_j)^{-0.5} \tag{2}$$

This descriptor was introduced by Kier and Hall in 1976 [22]. It effectively and simply represents the chemical structure and is readily calculated according to Eq. (2), where δ_i and δ_j are the values of the so-called atomic connectivity (where *i, j* pairs correspond to the connected atoms only). The atomic connectivity for the *i*-th atom in the molecular skeleton is defined as follows

$$\delta_i = \frac{Z_i - H_i}{Z_i - Z_i^v - 1} \tag{3}$$

where Z_i is the total number of electrons in the *i*-th atom, Z_i^v is the number of valence electrons and H_i is number of the hydrogens directly attached to the *i*-th atom.

The second descriptor is the *Total molecular two-center resonance energy divided by number of atoms, (E_R/N)*. This is a quantum-chemical descriptor calculated by the Ampac program from

$$E_R(tot) = 1/2 \sum_A \sum_{B \neq A} E_R(AB) \tag{4}$$

and

$$E_R(AB) = \sum_{\mu \in A} \sum_{\nu \in B} P_{\mu\nu} \beta_{\mu\nu} \tag{5}$$

where μ and ν are the atomic orbitals centered on atoms A and B, respectively. $P_{\mu\nu}$ are the density matrix elements and $\beta_{\mu\nu}$ are the one-electron two-center resonance integrals

$$\beta_{\mu\nu} = \left\langle \mu_A \left| -\frac{1}{2}\Delta - V_A - V_B \right| \nu_B \right\rangle \tag{6}$$

calculated in the atomic basis. The integral $\beta_{\mu\nu}$ represents the energy of a single electron moving in the Coulomb field of nuclei A and B described by the two-center density distribution $\mu_A \nu_B$.

Further proceeding to the three-parameter equation adds to the model the *Maximum net atomic charge for a C atom (Gaussian NBO).* It is also quantum-chemically calculated descriptor, which origins from the molecular charge distribution obtained *ab initio* using Gaussian program from the natural bond orbital population analysis [34]. This descriptor accounts for the polar interactions between molecules and may be employed as a measure of weak intermolecular interactions.

The fourth descriptor added to the previous model is the *Average valency of a C atom* that encodes features responsible for reactivity of the molecules. This descriptor was computed in the frames of the semiempirical population analysis performed by means of the Ampac program and can be easily extracted from its output. It is calculated as an average value of the quantum chemical valencies of all the carbon atoms in the molecule.

The correlation between the experimental data and the boiling points calculated using the final four-descriptor equation is presented graphically in Fig. 3. The small number of outliers shows that it can be used, with considerable confidence, for the prediction of the boiling point of ketones.

Table 2 summarizes the results of the best four-parameter QSPR model derived for boiling points of 202 ketones. The second column of this table contains all the molecular descriptors involved in this correlation, and in the third column there are the respective values of the linear coefficients of Eq. (1) along with their mean square errors. Finally, the fourth column contains the Student *t*-test values for the correlation coefficients, which reflect the significance of each parameter in the model. According to this test the most important descriptors of this model are *Total molecular two-center resonance energy/number of atoms* and *Kier&Hall index (order1)*.

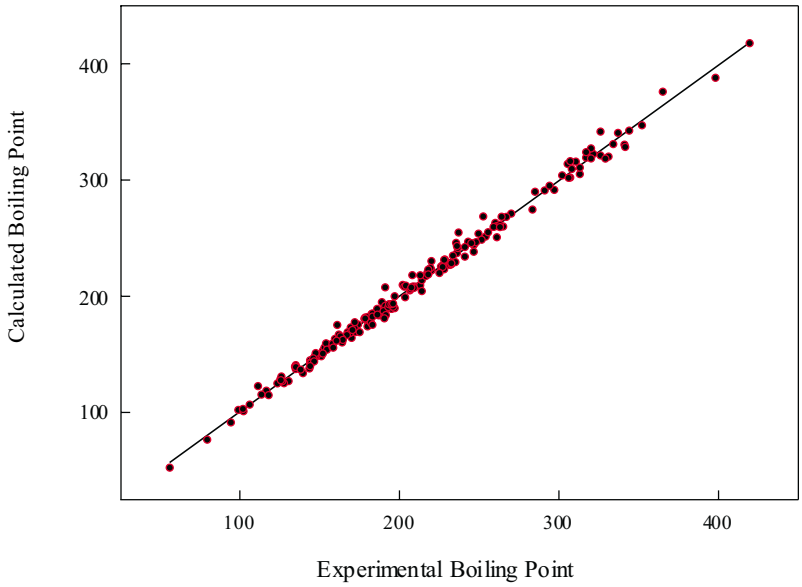

Figure 3. The experimental boiling point for 202 ketones *vs.* those calculated by the four-parameter correlation equation, $R^2 = 0.9907$, F = 5231.

Table 2. The parameters of the best four-descriptor correlation model, Eq. (1).

i	Descriptor, d_i	$a_i \pm \Delta a_i$	*t*-test
0	Intercept	–6074.3±610.1	–9.96
1	Kier&Hall index (order1)	35.4±0.73	48.90
2	Total molecular two-center resonance energy/number of atoms	–27.7±0.44	–62.56
3	Maximum net atomic charge for a C atom (Gaussian NBO)	–459.1±44.1	–10.43
4	Average valency of a C atom	1550.4±154.5	10.04

It is worth noting that in the present study the diverse set of ketones was treated as a whole, and no subsets, like aromatic, aliphatic, or cyclic, were distinguished. The Principal Component Analysis run for the final model exhibited no clustering, which strengthens the predictive power of the model for ketones regardless their structure.

SUMMARY

A Quantitative-Structure Property Relationship model was derived to study the boiling point of ketones. Innovation of the methodology presented is the combination of standard constitutional, topological, geometrical, and electrostatic descriptors with quantum-chemically calculated descriptors. They enabled obtaining the correlation model involving only four descriptors with high correlation factors $R^2 = 0.9907$ and $F = 5231$. The obtained QSPR equation can be applied to the prediction of boiling points of different groups of ketones, like aliphatic, aromatic and cyclic. The descriptors included in the correlation reflect the dependence of the boiling point on the chemical structure of the molecule, intermolecular dipol-dipol interaction and association forces.

Acknowledgment

I am grateful to Prof. Alan R. Katritzky for permission to use the CODESSA program. I am also thankful to Prof. M. Szafran for his encouragements and helpful discussions.

REFERENCES

1. Rechsteiner C.E., *Handbook of Chemical Property Estimation Methods,* Lyman W.J., Reehl W.F., Rosenblatt D.H., Eds.; McGrowHill: NY, Chapter 12 (1982).
2. Wold S. and Sjöström M., *Chemometr. Intell. Lab. Syst.*, **44**, 3 (1998).
3. Katritzky A.R., Lobanov V.S. and Karelson M., *Chem. Soc. Rev.,* **24**, 279 (1995).
4. Wiener H., *J. Am. Chem. Soc.*, **69**, 17 (1947).
5. Stanton D.T., Jurs P.C. and Hicks M.G., *J. Chem. Inf. Comput. Sci.*, **31**, 301 (1991).
6. Stanton D.T., Egolf L.M., Jurs P.C. and Hicks M.G., *J. Chem. Inf. Comput. Sci.,* **32**, 306 (1992).
7. Murugan R., Grendze M.P., Toomey J.E., Katritzky A.R., Karelson M., Lobanov V.S. and Rachwal P., *CHEMTECH*, **24**, 17 (1994).
8. Katritzky A.R., Lobanov V.S., Karelson M., Murugan R., Grendze M.P. and Toomey J.E., *Rev. Roum. Chim.*, **41**, 851 (1996).
9. Egolf L.M. and Jurs P.C., *J. Chem. Inf. Comput. Sci.*, **33**, 616 (1993).
10. Le T.D. and Weeres J.G., *J. Phys. Chem.*, **99**, 6739 (1995).
11. Espinosa G., Yaffe D., Cohen Y., Arenas A. and Giralt F., *J. Chem. Inf. Comput. Sci.,* **40**, 859 (2000).
12. Katritzky A.R., Mu L., Lobanov V.S. and Karelson M., *J. Phys. Chem,* **100**, 10400 (1996).
13. Katritzky A.R., Lobanov V.S. and Karelson M., *J. Chem. Inf. Comput. Sci.*, **38**, 28 (1998).
14. Stanton D.T., *J. Chem. Inf. Comput. Sci.,* **40**, 81 (2000).
15. Katritzky A.R., Lobanov V.S. and Karelson M., CODESSA, Reference Manual, University of Florida 1994.
16. Karelson M., Maran U., Wang Y. and Katritzky A.R., *Collect. Czech. Chem. Commun.*, **64**, 1551 (1999).
17. Carlson R., Prochazka M.P. and Lundstedt T., *Acta Chem. Scand.*, **B42**, 145 (1988).
18. Dewar M.J.S., Zoebisch E.G., Healy E.F. and Stewart J.J.P., *J. Am. Chem. Soc.,* **107**, 3902 (1985).
19. Gaussian 98, Revision A.10, Frisch M.J., Trucks G.W., Schlegel H.B., Scuseria G.E., Robb M.A., Cheeseman J.R., Zakrzewski V.G., Montgomery J.A., Jr., Stratmann R.E., Burant J.C., Dapprich S., Millam J.M., Daniels A.D., Kudin K.N., Strain M.C., Farkas O., Tomasi J., Barone V., Cossi M., Cammi R., Mennucci B., Pomelli C., Adamo C., Clifford S., Ochterski J., Petersson G.A., Ayala P.Y., Cui Q., Morokuma K., Salvador P., Dannenberg J.J., Malick D.K., Rabuck A.D., Raghavachari K., Foresman J.B., Cioslowski J., Ortiz J.V., Baboul A.G., Stefanov B.B., Liu G., Liashenko A., Piskorz P., Komaromi I., Gomperts R., Martin R.L., Fox D.J., Keith T., Al-Laham M.A., Peng C.Y., Nanayakkara A., Challacombe M., Gill P.M. W., Johnson B., Chen W., Wong M.W., Andres J.L., Gonzalez C., Head-Gordon M., Replogle E.S. and Pople J.A., Gaussian, Inc., Pittsburgh PA, 2001.

20. Szafran M., *Wiad. Chem.,* **47**, 477 (1993).
21. Randic M., *J. Am. Chem. Soc.*, **97**, 6609 (1975).
22. Kier L.B. and Hall L.H., *Molecular Connectivity in Chemistry and Drug Research,* Academic Press, NY, (1976).
23. Kier L.B. and Hall L.H., *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth (1986).
24. *Handbook of Chemistry and Physics*, CRC Press, Cleveland OH, 1974, p. 112.
25. Rohrbaugh R.H. and Jurs P.C., *Anal. Chim. Acta.*, **199**, 99 (1987).
26. Stouch T.R. and Jurs P.C., *J. Chem. Inf. Comput. Sci.*, **26**, 4 (1986).
27. Pearlman R.S., *Physical Chemical Properties of Drugs*, Eds.: Yalkowsky S.H., Sinkula A.A. and Valvani S.C., Marcel Dekker Inc., NY, 1980, p. 321.
28. Zefirov N.S., Kirpichenok M.A., Izmailov F.F. and Trofimov M.I., *Dokl. Akad. Nauk SSSR*, **296**, 883 (1987).
29. Osmialowski K., Halkiewicz J., Radecki A. and Kaliszan R., *J. Chromatogr.*, **346**, 53 (1985).
30. Osmialowski K., Halkiewicz J. and Kaliszan R., *J. Chromatogr.*, **361**, 63 (1986).
31. Stanton D.T. and Jurs P.C., *Anal. Chem.*, **62**, 2323 (1990).
32. Stanton D.T., Egolf L.M., Jurs P.C. and Hicks M.G., *J. Chem. Inf. Comput. Sci.*, **32**, 306 (1992).
33. Karelson M., Lobanov V.S. and Katritzky A.R., *Chem. Rev.*, **96**, 1027 (1996).
34. Reed A.E., Curtiss L.A. and Weinhold F., *Chem. Rev.*, **88**, 899 (1988).