Full Articles

Fragmental descriptors in QSPR: application to molecular polarizability calculations

N. I. Zhokhova, I. I. Baskin, V. A. Palyulin, A. N. Zefirov, and N. S. Zefirov★

Department of Chemistry, M. V. Lomonosov Moscow State University,
Leninskie Gory, 119992 Moscow, Russian Federation.
Fax: +7 (095) 939 0290. E-mail: zhokhova@org.chem.msu.ru; zefirov@org.chem.msu.ru

Applicability of the fragmental approach developed in the framework of the QSPR methodology to prediction of the molecular polarizability of various classes of organic compounds is demonstrated. The model proposed allows reliable prediction of the molecular polarizability of organic compounds based on their chemical composition and a set of fragmental descriptors, which characterize the multiple and aromatic bonds as well as fused aromatic systems.

Key words: QSPR methodology, molecular polarizability, organic compounds, fragmental approach, fragmental descriptors.

The polarizability of a molecule (α) is one of the most significant electrical properties, which characterizes the ability of the electronic system to be distorted by the external electric field.[1,2] It is defined as the coefficient of proportionality between the strength of an applied electric field (E) and the magnitude of the induced dipole moment (μ_{ind}) using the equation

$$\mu_{\mathrm{ind}} = \alpha E.$$

In principle, the polarizability of molecules is determined by the strength of the attractive interaction between electrons and atomic nuclei. Usually, if a molecule contains a few electrons, its polarizability is lower than that of the molecule containing atoms with a larger number of electrons and a more diffuse electron distribution. Experimentally, polarizability is determined from the molar refraction ($\mathrm{MR_D}$) values, which are calculated using the refractive index n_{D}, the density ρ, and the molecular weight M by the Lorentz—Lorenz relation[1—3]

$$\mathrm{MR_D} = \left(\frac{n_{\mathrm{D}}^2 - 1}{n_{\mathrm{D}}^2 + 2}\right)\frac{M}{\rho} = \frac{4}{3}\pi N_0\alpha,$$

where N_0 is the Avogadro constant. If the parameter $\mathrm{MR_D}$ is expressed in cm^3 and α in Å^3, one gets $\alpha = 0.3964\mathrm{MR_D}$.

The molar refraction is an additive property. A number of additive models for calculations of this quantity using both the atomic increments and the bond increments have been proposed.[1—3] In other words, this means that the various atoms or functional groups can be assigned refraction values (increments) that are constant for a variety of molecules and that the sum of these values is

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 5, pp. 1005—1009, May, 2003.

the molar refraction. Because of this the molar refraction was of crucial importance for structure assignment in chemical research in the pre-spectroscopic era.[1–3] It seems reasonable to extend the additivity approach to the polarizability.

A large number of more or less sophisticated methods of different accuracy was proposed for calculating the molecular polarizabilities.[4–7] In particular, the polarizability tensors were ascribed to particular bonds and functional groups in accordance with the hypothesis that component summation of the group tensors gives the molecular polarizability tensor.[5] A semiempirical method developed on this basis to calculate the components of the molecular polarizability tensor gave a standard deviation of 3.5% for the set containing 120 structures.[6,7]

However, the results obtained in Ref. 8 appeared to be the most surprising. It was shown that a simple ten-descriptor model using the atomic polarizabilities for ten elements allows rather accurate calculations of the molecular polarizabilities without considering any other structural parameters. The training set contained 340 compounds and the test set contained 86 compounds (hereafter, Database *2*);[8] the QSPR equation is as follows

$$\begin{aligned}\alpha_{\text{calc}} = {} & 0.32 + 1.51N_{\text{C}} + 0.17N_{\text{H}} + 0.57N_{\text{O}} + 1.03N_{\text{N}} + \\ & + 2.99N_{\text{S}} + 2.48N_{\text{P}} + 0.22N_{\text{F}} + 2.16N_{\text{Cl}} + \\ & + 3.29N_{\text{Br}} + 5.45N_{\text{I}}, \qquad (1)\end{aligned}$$

$n = 340$, $r^2 = 0.9989$, $s = 0.33$, $F = 20338$,

where N is the number of atoms of the corresponding element, n is the number of compounds in the set, r is the correlation coefficient, s is the standard deviation, and F is the Fisher test.

It should be emphasized that the results obtained using quantum-chemical methods (*e.g.*, MINDO, AM1, PM3, and even DFT) are worse than those obtained by the additive approach. Therefore, it was of considerable interest to check the additive Eq. (1) for versatility using a larger set.

Recently, the QSPR methodology has been successfully employed for calculations of various physicochemical characteristics. Therefore, it was of interest to use this approach in the molecular polarizability calculations. The results of the molar refraction calculations using the topological indices[9,10] and the neural networks[8,11] are available.

Earlier,[12–17] we have developed the FRAGMENT subroutine capable of generating sets of structural fragments, namely, chains one to nine atoms long, rings (from three-membered to six-membered ones), and several types of branched fragments. What is more, each atom in a fragment is encoded depending on its type, the number of attached H atoms, and the bonding environment, thus providing some flexibility in taking into account of the multiple bonds, functional groups, heteroatoms, *etc*. The FRAGMENT subroutine was successfully employed in the QSAR/QSPR program package EMMA[17–25] and the NASAWIN neural network program.[26–28] A modified version of the FRAGMENT subroutine is capable of generating a larger number of fragment types and incorporates tools for more flexible classification of atoms.[17,29]

In our previous QSAR/QSPR studies we have widely used the fragmental descriptors for both the QSPR (chromatographic retention indices and boiling points of various classes of chemical compounds) and QSAR predictions.[17,25] A point of crucial importance should be emphasized. If a set of compounds is sufficiently large to construct a statistically significant model, any topological index can be replaced by a set of fragmental descriptors.[17,30] The advantages of using the fragmental descriptors are also "transparency" and ready interpretability of the results of QSAR/QSPR studies. In this work we report the results of the QSPR treatment of molecular polarizability using structural descriptors.

Calculation Procedure

QSPR calculations were carried out using the QSAR/QSPR programs EMMA[17–25] and NASAWIN.[26–28] Databases were created using the MEOW and BASTET programs* developed for the manual input and sorting of the QSAR/QSPR structural databases and search for duplicates. These programs are convenient for joint use with the QSAR program packages EMMA and NASAWIN. The structural databases can be automatically converted into various file formats including the sdf format.

Results and Discussion

Two databases were initially created to be used in this study. First, using the results reported in Refs. 6 and 7, we created the Database *1* containing 293 structures corresponding to various classes of organic compounds and to some inorganic substances (*e.g.*, H_2, O_2, N_2, N_2O, CO, SO_2, H_2S, H_2O, NH_3, Cl_2, *etc*.). Then, we created the Database *2* which included the structures of 426 compounds containing the C, H, O, N, S, P, F, Cl, Br, and I atoms (cyclic and acyclic non-aromatic hydrocarbons, aromatic hydrocarbons, their halo derivatives including perfluorinated compounds, as well as alcohols, phenols, ethers and esters, aldehydes, ketones, carboxylic acids, amines, nitriles, nitro derivatives, amides, and suflur- and phosphorus-containing compounds) using the results reported in Ref. 8. In processing the published data[8] using the BASTET program we found that the structures Nos. 29 and 226, 158 and 200, and 236 and 260 are duplicates. In a private communication the authors of Ref. 8 confirmed

* The MEOW and BASTET programs are available on request.

that the first two pairs of compounds are duplicates and pointed out that the structure No. 236 is correct ($O{=}P(NMe_2)_3$), whereas the structure No. 260 corresponds to a formula of methylphosphonic acid bis(dimethylamide) ($O{=}P(Me)(NMe_2)_2$). Therefore, the training set for the Database *2* contained a total of 338 rather than 340 structures.

Using the BASTET program, we also created a joint database (Database *3*), which combined the Databases *1* and *2* and contained a total of 613 structures after exclusion of duplicates. If the experimental data in the Database *1* were different from the corresponding data in the Database *2* (a few cases with mostly insignificant differences), the structures from the Database *2* were included in the Database *3*.

First of all we performed a validity check of the additive scheme proposed earlier[3] using the extended Database *3*. After calculating the averaged atomic polarizabilities using the additive model[8] we obtained a very good correlation with the experimental values (Fig. 1):

$$\alpha_{exp} = 1.035\alpha_{calc} - 0.355,$$

$$n = 613, r^2 = 0.9898, s = 0.613.$$

Thus, despite a considerable extension of the working database, the additive model allows highly accurate calculations of the polarizability, though the standard deviation in the case of the Database *3* increased to some extent ($s = 0.613$, *cf.* $s = 0.33$ for the calculations with the Database *2*). However, by introducing additional descriptors accounting for the multiple bonds and aromatic systems it is possible to improve the computational scheme and appreciably reduce the error in calculations with the Database *3* (see below).

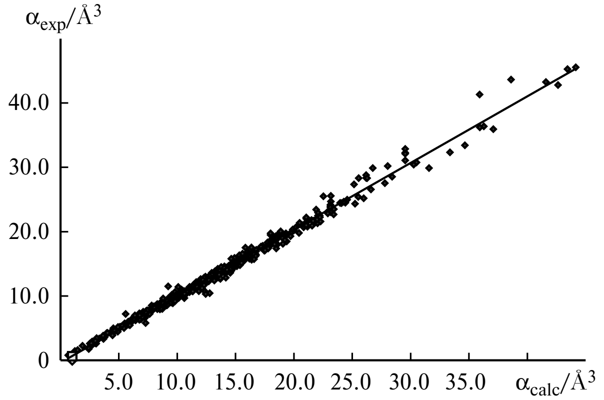

Fig. 1. Correlation between the calculated (α_{calc}) and experimental (α_{exp}) molecular polarizability values obtained for the Database *3* using the additive model (Eq. (1)).

Let us examine the results of the QSPR study in more detail. The simplest fragmental descriptors were calculated using the ELEM block. The restrictions imposed on the operation of the FRAGMENT subroutine were as follows: (i) structural fragments contained one to five atoms (if the EMMA program package was employed) or one to six atoms (if the NASAWIN program was used), (ii) fragmental descriptors were selected both manually and using step-wise regression, and (iii) fragments containing the smallest number of atoms were selected from the groups of the mutually correlating descriptors. The training set and the test set were chosen using the following procedure. First, the test sets for the Databases *1*

Table 1. Statistical parameters (n, r, s, F) of the best QSPR fragmental descriptor models for calculating the molecular polarizability values

Model	Training set						Test set		
	database	number of descriptors	n	r^2	s	F	set	n	RMS[a]
1	*1*	13[b]	190	0.9951	0.69	2537	Database *2*	423	0.66
2	*1*	14[c]	264	0.9951	0.58	3382	10% of compounds in the set[d]	29	0.58
3	*2*	13	423	0.9958	0.31	6854	Database *1*	190	0.98
4	*2*	10[e]	338	0.9943	0.35	5206	Compounds 339—423	85	0.35
5	*2*	13	338	0.9954	0.31	5048	The same	85	0.31
6	*2*	14	338	0.9955	0.31	4755	»	85	0.31
7	*3*	10	552	0.9936	0.53	7640	10% of compounds in the set[d]	61	0.69
8	*3*	13	552	0.9947	0.49	7191	The same	61	0.70
9	*3*	14	552	0.9967	0.38	10931	»	61	0.75

[a] Root-mean-square deviation.

[b] Descriptors characterizing the number of the C, H, N, O, S, P, F, Cl, Br, and I atoms and the numbers of the multiple and aromatic bonds in the molecule.

[c] Descriptors characterizing the number of the C, H, N, O, S, P, F, Cl, Br, and I atoms, the number of the multiple and aromatic bonds, and the number of junctions in the aromatic system of the molecule.

[d] Each tenth compound.

[e] Descriptors characterizing the number of the C, H, N, O, S, P, F, Cl, Br, and I atoms in the molecule.

and *2* were the same as those used in the original studies.[6—8] Second, the Databases *1* and *2* were used in turn as the training set and as the test set. In this case duplicates of the compounds from the Database *2* were excluded from the Database *1*, so that the total number of compounds in the Database *1* was 190. Third, a QSPR study was carried out with the Database *3* as the training set and with each tenth compound from this database as the test set. More than 150 linear regression models were constructed. The statistical parameters of the best models are listed in Table 1.

As can be seen in Table 1, the use of the fragmental approach permits construction of models with high predictive power for all the databases used. In particular, in all cases the squared correlation coefficient is at least 0.9936 while the standard deviation is at most 0.69.

Among all the sets studied, the lowest correlation coefficients were obtained for the ten-descriptor models using the number of atoms of different elements in the molecule (see Table 1, models 4 and 7). This is not surprising since the structural fragments were selected by the smallest size and in these models we obtained, in essence, the simple atom-by-atom sum. In other words, in the simplest case the use of the fragmental QSPR approach automatically reproduced the additive scheme.

The model 7 was constructed using the most representative set (Database *3*). In this case the largest errors were found for 1,2:5,6-dibenzoanthracene (error is 3.72 Å^3), coronene (3.05 Å^3), 2,3:4,5-dibenzophenazine (2.6 Å^3), dodecahydrotriphenylene (2.5 Å^3), acridine (2.1 Å^3), and penta-1,4-diene (2.1 Å^3). Examination of this list makes possible the following assumption. Since the molecules of these compounds contain a large number of the multiple and aromatic bonds and the aromatic systems in many molecules are fused, one can expect that introduction of additional descriptors to account for the number of such bonds as well as the number of fused aromatic rings will improve the QSPR models. In other words, increasing the number of the descriptors responsible for polyunsaturation or aromaticity of the condensed systems in a targeted manner allows one to expect improvement of the computational scheme (1) and the models 4 and 7 (see Table 1).

This assumption was tested experimentally. For instance, the improved thirteen-descriptor models 1, 3, 5, and 8 (see Table 1) have better statistical characteristics compared to the models examined above (*cf.* the models 4 and 5 and the models 7 and 8). Moreover, introduction of a fragmental descriptor to account for the number of junctions in the aromatic systems ($C_{Ar}(C_{Ar})_2$) improves the predictive power of the models for the training set (see Table 1, the models 7, 8, and 9), though does not improve it for the test set. This is probably due to the lack of statistically representative data sets for the polycondensed systems in the test set.

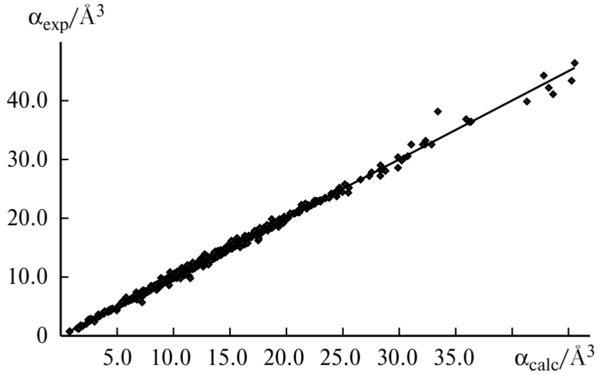

Fig. 2. Correlation between the calculated (α_{calc}) and experimental (α_{exp}) molecular polarizability values obtained for the Database *3* using the model 9 (Eq. (2)).

Figure 2 presents a correlation between the calculated and experimental values of the molecular polarizabilities, obtained for the best (based on the correlation coefficient, standard deviation, and the Fisher test) model 9 constructed using the largest set (Database *3*). The QSPR equation for the polarizability calculations has the form

$$\alpha_{calc} = -0.04 + 1.08f_1 + 0.38f_2 + 0.92f_3 + 0.61f_4 + \\ + 3.04f_5 + 2.18f_6 + 0.44f_7 + 2.34f_8 + 3.35f_9 + \\ + 5.49f_{10} + 0.38f_{11} + 0.15f_{12} + 0.34f_{13} + 0.36f_{14}, \qquad (2)$$

$n = 552$, $r^2 = 0.9967$, $s = 0.38$, $F = 10931$,

where $f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}$ is the number of C, H, N, O, S, P, F, Cl, Br, and I atoms, respectively; f_{11} is the number of triple bonds in the molecule; f_{12} is the number of double bonds in the molecule; f_{13} is the number of aromatic bonds; and f_{14} is the number of atoms in the ring junctions in the aromatic system ($C_{Ar}(C_{Ar})_2$). For this model the largest errors in predicting the polarizability were also obtained for the condensed polyaromatic compounds, coronene (error is 2.5 Å^3) and 2,3:4,5-dibenzophenazine (4.7 Å^3). Therefore, though introduction of additional descriptors improves the statistical characteristics, it is not always sufficient for more accurate description of such complex systems.

Thus, in the framework of the QSPR methodology we obtained the relations which permit highly accurate prediction of the molecular polarizability of different classes of organic compounds based on their elemental composition and on the number of fragments characterizing the multiple and aromatic bonds and those responsible for the condensed aromatic systems.

References

1. C. D. Neniţescu, *Chimie Organică*, Editure Tehnica, Bucuresti, 1960.

2. S. S. Batsanov, *Strukturnaya refraktometriya* [*Structural Refractometry*], Vysshaya shkola, Moscow, 1976, 302 pp. (in Russian).
3. S. S. Batsanov, *Eksperimental´nye osnovy strukturnoi khimii* [*Experimental Foundations of Structural Chemistry*], Izd-vo standartov, Moscow, 1986, Ch. 3.2 and 3.3 (in Russian).
4. J. M. Stout and C. E. Dykstra, *J. Am. Chem. Soc.*, 1995, **117**, 5127.
5. J. Applequist, J. R. Carl, and K. Fung, *J. Am. Chem. Soc.*, 1972, **94**, 2952.
6. K. J. Miller, *J. Am. Chem. Soc.*, 1990, **112**, 8533.
7. K. J. Miller, *J. Am. Chem. Soc.*, 1990, **112**, 8543.
8. R. Bosque and J. Sales, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 1154.
9. A. R. Katritzky, S. Sild, and M. Karelson, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 840.
10. A. R. Katritzky, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 1171.
11. S. Liu, R. Zhang, M. Liu, and Z. Hu, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1146.
12. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Tez. dokl. Mezhvuz. konf.* "*Molekulyarnye grafy v khimicheskikh issledovaniyakh*" [*Abstrs. Higher School Conf.* "*Molecular Graphs in Chemical Research*"], Kalinin, 1990, 5 (in Russian).
13. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Tez. dokl. 1-i Vsesoyuz. konf. po teoreticheskoi organicheskoi khimii* [*Abstrs. 1st All-Union Conf. on Theoretical Organic Chemistry*], Volgograd, 1991, 557 (in Russian).
14. V. A. Palyulin, I. I. Baskin, D. E. Petelin, and N. S. Zefirov, *Abstrs. 10th Eur. Symp. on Structure—Activity Relationships: QSAR and Molecular Modelling*, Barcelona, 1994, B257.
15. V. A. Palyulin, I. I. Baskin, D. E. Petelin, and N. S. Zefirov, in *QSAR and Molecular Modelling: Concepts*, *Computational Tools and Biological Application*s, Eds. F. Sanz, J. Giraldo, and F. Manaut, Prous Science Publishers, Barcelona, 1995, 51.
16. V. A. Palyulin, E. V. Radchenko, I. I. Baskin, A. Yu. Zotov, and N. S. Zefirov, *Abstrs. 11th Eur. Symp. on QSAR: Computer Assisted Lead Finding and Optimization*, Lausanne, 1996, 31A.
17. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2002, **41**, 1112.
18. D. E. Petelin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1992, **324**, 1019 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
19. T. S. Pivina, D. V. Sukhachev, and L. K. Maslova, *Dokl. Akad. Nauk*, 1993, **330**, 468 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
20. D. V. Sukhachev, T. S. Pivina, V. A. Shlyapochnikov, E. A. Petrov, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 1993, **328**, 188 [*Dokl. Chem.*, 1993 (Engl. Transl.)].
21. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, and S. I. Zeman, *Izv. Akad. Nauk*, *Ser. Khim.*, 1995, 1653 [*Russ. Chem. Bull.*, 1995, **44**, 1585 (Engl. Transl.)].
22. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, and N. S. Zefirov, *Izv. Akad. Nauk*, *Ser. Khim.*, 1995, 1657 [*Russ. Chem. Bull.*, 1995, **44**, 1589 (Engl. Transl.)].
23. D. V. Sukhachev, T. S. Pivina, N. I. Zhokhova, N. S. Zefirov, and S. I. Zeman, *Izv. Akad. Nauk*, *Ser. Khim.*, 1995, 1661 [*Russ. Chem. Bull.*, 1995, **44**, 1594 (Engl. Transl.)].
24. N. S. Zefirov and V. A. Palyulin, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1022.
25. N. S. Zefirov, V. A. Petelin, V. A. Palyulin, and J. McFarland, *Dokl. Akad. Nauk*, 1992, **327**, 504 [*Dokl. Chem.*, 1992 (Engl. Transl.)].
26. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, in *QSAR and Molecular Modelling: Concepts*, *Computational Tools and Biological Application*s, Eds. F. Sanz, J. Giraldo, and F. Manaut, Prous Science Publishers, Barcelona, 1995, 30.
27. N. M. Halberstam, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Proc. Int. Symp. CACR-96*, Moscow, 1996, 37.
28. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 715.
29. N. V. Artemenko, I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *Dokl. Akad. Nauk*, 2001, **381**, 203 [*Dokl. Chem.*, 2001 (Engl. Transl.)].
30. I. I. Baskin, M. I. Skvortsova, I. V. Stankevich, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 527.

Received November 19, 2002