Fragmental Approach in QSPR

Nikolai S. Zefirov* and Vladimir A. Palyulin

Department of Chemistry, Moscow State University, Moscow 119992, Russia

Received February 25, 2002

Methodological problems of using *fragmental descriptors* for construction of QSAR/QSPR equations are considered, and the main achievements in this field are summarized and discussed. If a structure−property data set is sufficiently large to allow building statistically significant models, then *any topological index can be replaced with a set of fragmental descriptors*. Several examples of using the fragmental approach for predicting retention indices and the normal boiling points of organic compounds are considered. Advantages of using fragmental descriptors, namely a "transparency" and interpretability of QSAR/QSPR models, are exemplified.

INTRODUCTION

"The basis of the science of organic chemistry is the *structural theory*. It is the basis upon which millions of facts about hundreds of thousands of individual compounds have been brought together and arranged in a systematic way. It is the basis upon which these facts can best be accounted and understood".[1] A heart of structural theory is a structural formula. Knowledge of the formula permits one to grasp the essential knowledge about chemical behavior of chemical species.

In the framework of structural theory, many additional important notions have been introduced into chemistry, such as, for instance, *homology* and *functionality*. Functional group is a fragment of structural formula, which is transferable and, moreover, is a carrier of some particular set of transferable properties. Thus, it was a first basis for rationalization of properties depending on molecular weight and on particular atomic sets and bindings. Indeed, a property of an organic compound may be considered as dependent on the presence of some functional groups (or "substituents"), each of them making a contribution into it.

Moreover, this approach permits to make *quantitative* considerations: in the framework of structural theory it is very natural to use *additive schemes* for treatment of properties. In this case, some property is considered as partitioned into atomic or bond contributions

$$A = \sum_{1}^{N} A_i \cdot n_i \qquad (1)$$

where A is a value of some property for a compound, n_i is the occurrence number of a structural fragment (either atom or bond, depending on scheme) of type i in this compound, A_i is the contribution of the corresponding fragment into the property, and N is the overall number of fragment types. Probably the best additive schemes, both atomic[2] and bonding,[3] were elaborated for molecular refraction, and they were extremely important for structural elucidation in pre-spectroscopic era. Let us also point out well-known additive schemes for parachore,[4] formation or combustion enthalpies[5] (often using a notion "bond energies"), lipophilicity,[6−8] etc.

It was early understood that simple atom or bond additivity does not permit the satisfactory explanation of many types of properties, that, in turn, gave rise to necessity to apply more sophisticated schemes accounting for atoms or bonds surroundings. One of the most developed systematics of additive schemes with introduction of extensive systems of subtypes was made by Tatevsky and co-workers.[9]

In essence, analogous logic was used in Hammett approach to the problem of quantitative structure−reactivity relationships, where the property of some "core fragment" is treated as independently modified by "substituents", having specific characteristic values. Further development of these ideas by Hansch[6,7] led to creation of QSAR[10] (Quantitative Structure−Activity Relationships) as a general avenue of modern approaches to analysis of properties, including biological ones. Appearance of QSAR/QSPR as well as application of graph theory to the structural problems had a tremendous impact on further development of fragmental approaches.

Probably Smolensky[11,12] was the first who used subgraph approach for calculation of physicochemical properties. The property was considered as a linear function of many variables, each of them being the number of particular subgraphs (chains) in a molecular graph. The contributions of one-atomic fragments were considered as the main ones, those of two-atomic fragments were considered as correction terms for the first surrounding, three-atomic fragments were responsible for the second surrounding, etc. The fruitfulness of this idea was proven by further investigations (vide infra).

Broadly speaking, all fragmental methods are based on the general formula (2)

$$A = A_0 + \sum_{1}^{N} A_i \cdot n_i \qquad (2)$$

where the total property A is represented as a sum of properties of fragments added to some constant A_0.[13]

Moreover, the idea of characterizing the molecular structure by a series of subgraphs of increasing size has support

* Corresponding author phone: (7)095-9391620; fax: (7)095-9390290; e-mail: zefirov@org.chem.msu.su and vap@org.chem.msu.su.

10.1021/ci020010e CCC: $22.00
Published on Web 07/31/2002

from graph-theoretical point of view. It was recently proven[14] that any molecular graph invariant (that is any topological index, TI) can be represented as (1) a linear combination of occurrence numbers of some substructures (fragments), both connected and disconnected, or (2) a polynomial of occurrence numbers of connected substructures of corresponding molecular graph. In principle, this means that *if a structure−property data set is sufficiently large to allow building statistically significant models, then any topological index can be replaced with a set of substructural descriptors.*[14]

The ramifications of further development differ by definitions, ways of search, and strategies of accounting structural fragments. One of the most known is Free-Wilson model, which is based on a multiple regression analysis that uses "indicator variables" (that indicate the presence of fragments, which are usually the "substituents" in a "core fragment") as descriptors of molecules.[15] The close similarity with Hammett approach is evident. In accordance with that, the fragments are usually simple substituents such as small alkyls, halogens, nitrogen-, and oxygen-containing groups, etc.[16] What is more, this methodology was also very successful in combination with graph theoretical topological indices or with some experimentally determined parameters (e.g. log P) even in very simplistic form, such as use of indicator variables, accounting numbers of atoms or bonds of particular types, total number of non-hydrogen atoms or some specific functions, etc.

On the other hand, the fragmental approach was transformed using graph theoretical approaches into count of "paths" of different lengths[17,18] and "fingerprints"[19] (or into more sophisticated calculations of "molecular walks"[20] or "detour indices"[21]), etc., which take into account in many cases just simple unlabeled subgraphs, often even not multigraphs. For the purpose of QSAR/QSPR all these fragmental descriptors are considered as variables, which are treated using various techniques such as linear regression (*LR*), multiple linear regression (*MLR*), partial least squares (*PLS*), artificial neural networks (*ANN*), etc.

For analysis of biological activity, a special language, *SSFN* (Substructure Superposition Fragment Notation), aimed to describe biologically important structural features was devised.[22] *SSFN* was advantageously applied for predicting the spectrum of biological activities by many authors;[23a−d] however, recently for this purpose another fragmental approach MNA ("Multilevel Neighborhoods of Atoms") had been developed.[23e,f]

We have to specially point out the approaches of Klopman (*CASE*-Approach)[24] and of Meylan-Howard (*AFC* - Atom/Fragment Contribution Method).[25] In both approaches, a molecular property is approximated by summation of local contributions from different fragments, while contribution from a fragment depends on its local neighborhood in a molecule. In Klopman's approach rather complex fragments, which are in some cases associated with the notions of biophores and biophobes, are generally used. In Meylan-Howard's approach, one-atomic fragments are predominantly used, while the complex nature of molecular structure is taken into account by means of numerous correction factors.

Probably the first use of *ANN* in combination with fragmental descriptors was demonstrated for the prediction of mutagenic activity[26] and in our work[27] for predicting physicochemical properties of alkanes. The *ANN* in combination with occurrence numbers of the simplest one-atomic fragments was used for predicting physicochemical properties and biological activity of organic compounds.[28] We will discuss this aspect of application of fragmental descriptors in future publications.

We should also point to a concept of molecular hologram, which is a vector representing the presence of various fragments in a chemical structure.[29] In the framework of this approach, a search for structure−property (biological activity) relationships of chemicals is carried out by means of *PLS* that allows processing of numerous mutually correlated descriptors. This technique makes it possible to interpret models by means of color-coding of fragments in dependence upon their contributions to molecular property (activity). Some other applications of fragmental descriptors see also in ref 30.

In 1990 we have developed the FRAGMENT program, which was able to generate extensive sets of fragments: chains (1−6 atoms), cycles (three−six-membered), and several types of branched fragments.[31] What is more, each atom in a fragment is coded depending on its type and neighborhood providing significant flexibility to account for heteroatoms, functionality, bond types, etc. The FRAGMENT was successfully incorporated into the program package EMMA[31] which, in turn, was extensively used for our QSAR/QSPR studies.[32−34] The FRAGMENT program was also incorporated into our neural network NASAWIN package.[35,36] Recently we have developed the modernized version of FRAGMENT program and the related software tools, in particular, for applications using neural networks (will be published).

The goal of this paper is to demonstrate the methodology and successful applicability of fragmental descriptors for QSPR purposes in comparison with topological indices.

QSPR STUDIES USING FRAGMENTAL DESCRIPTORS: RESULTS AND DISCUSSION

QSPR Results. To demonstrate the peculiarities of fragments as descriptors for QSPR studies we have used the data sets (DS) on retention indices in gas−liquid chromatography and boiling points. The relevance of these properties is quite obvious: there exist numerous works dealing with prediction of retention indices from boiling points and vice versa.[37] We used the following data for our purposes. First, we have used the data from the paper[38] containing the QSPR models for retention indices (RI) in gas−liquid chromatography of 50 alkylphenols (DS−RI; Table 1). Second, we have used four data sets for normal boiling points. One set represents the Balaban's et al. data (DS−NBP) of acyclic ethers, peroxides, acetals, and their sulfur analogues.[39] The second one is the DS of boiling points of 100 aliphatic alcohols (DS−BPROH), which was used to compare QSPR models generated by CODESSA vs models based on variable connectivity index.[40] The third data set is the very detailed collection of boiling points of a great variety of saturated hydrocarbons.[41a] This data set was a source of creation of different subsets; for example Trinajstić et al. used a subset of 180 hydrocarbons to investigate the application of different TIs, including novel distance-related ones.[21a] We also have used these data. The multiple linear regression QSPR calculations were made using software package EMMA.[31−33]

Table 1. Retention Indices of Alkylphenols[38] and the Values of Fragmental Descriptors[a]

no.	compound	RI	D1	D2	D3	D4	D5	D6	D7
1	phenol	1281	0	0	0	0	0	0	0
2	2-methylphenol	1354	1	0	1	0	0	0	0
3	3-methylphenol	1386	1	0	0	0	0	0	0
4	4-methylphenol	1385	1	0	0	0	0	0	0
5	2-ethylphenol	1430	2	0	1	1	0	1	0
6	3-ethylphenol	1483	2	0	0	1	0	2	0
7	4-ethylphenol	1473	2	0	0	1	0	2	0
8	2,3-dimethylphenol	1495	2	1	2	0	0	0	0
9	2,4-dimethylphenol	1456	2	0	1	0	0	0	0
10	2,5-dimethylphenol	1453	2	0	1	0	0	0	0
11	2,6-dimethylphenol	1416	2	0	2	0	0	0	0
12	3,5-dimethylphenol	1489	2	0	0	0	0	0	0
13	3,4-dimethylphenol	1530	2	1	1	0	0	0	0
14	4-isopropylphenol	1527	3	0	0	2	0	0	0
15	2-*n*-propylphenol	1502	3	0	1	1	1	1	1
16	3-*n*-propylphenol	1565	3	0	0	1	0	2	2
17	4-*n*-propylphenol	1563	3	0	0	1	0	2	2
18	2-ethyl-4-methylphenol	1523	3	0	1	1	0	1	0
19	2-ethyl-5-methylphenol	1529	3	0	1	1	0	1	0
20	2-ethyl-6-methylphenol	1485	3	0	2	1	0	1	0
21	3-ethyl-5-methylphenol	1581	3	0	0	1	0	2	0
22	4-ethyl-2-methylphenol	1539	3	0	1	1	0	2	0
23	4-ethyl-3-methylphenol	1608	3	1	1	1	0	1	0
24	2,3,4-trimethylphenol	1638	3	2	3	0	0	0	0
25	2,3,5-trimethylphenol	1593	3	1	2	0	0	0	0
26	2,3,6-trimethylphenol	1551	3	1	3	0	0	0	0
27	2,4,5-trimethylphenol	1593	3	1	2	0	0	0	0
28	3,4,5-trimethylphenol	1667	3	2	2	0	0	0	0
29	4-*sec*-butylphenol	1612	4	0	0	2	0	0	0
30	2-*n*-butylphenol	1600	4	0	1	1	1	1	1
31	3-*n*-butylphenol	1668	4	0	0	1	0	2	2
32	4-*n*-butylphenol	1661	4	0	0	1	0	2	2
33	2-methyl-4-*n*-propylphenol	1623	4	0	1	1	0	2	2
34	2-methyl-6-*n*-propylphenol	1553	4	0	2	1	1	1	1
35	3-methyl-6-*n*-propylphenol	1602	4	0	1	1	1	1	1
36	4-methyl-2-*n*-propylphenol	1593	4	0	1	1	1	1	1
37	2,4-diethylphenol	1602	4	0	1	2	0	3	0
38	2,5-diethylphenol	1624	4	0	1	2	0	3	0
39	3,4-diethylphenol	1682	4	1	1	2	0	2	0
40	2,3,4,5-tetramethylphenol	1782	4	3	4	0	0	0	0
41	2,3,4,6-tetramethylphenol	1690	4	2	4	0	0	0	0
42	2,3,5,6-tetramethylphenol	1683	4	2	4	0	0	0	0
43	2-ethyl-4,5-dimethylphenol	1656	4	1	2	1	0	1	0
44	2-*n*-pentylphenol	1700	5	0	1	1	1	1	1
45	4-*n*-pentylphenol	1765	5	0	0	1	0	2	2
46	4-*tert*-pentylphenol	1703	5	0	0	3	0	0	0
47	2-ethyl-5-*n*-propylphenol	1706	5	0	1	2	0	3	2
48	2-*n*-hexylphenol	1800	6	0	1	1	1	1	1
49	4-*n*-hexylphenol	1871	6	0	0	1	0	2	2
50	3-*n*-butyl-6-ethylphenol	1807	6	0	1	2	0	3	2

[a] **D1** C_{sp3}; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $RC_{Ar}\div C_{Ar}R$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $CH_2-CH_2-C_{Ar}\div C_{Ar}R$; **D6** $CH_2-C_{Ar}\div C_{Ar}H$; **D7** $CH_2-CH_2-C_{Ar}\div C_{Ar}H$. R means here that aromatic carbon has a substituent.

For DS−RI[38] (Tables 1 and 2) we have constructed five QSPR models using fragmental descriptors: (1) one model for the whole set, which is the following: RI = 1288 + 100.8**D1** +79.9**D2** −39.7**D3** −33.1**D4** −24.1**D5** + 9.5**D6** −5.8**D7**, $N = 50$, $R^2 = 0.9981$, $s = 5.8$, $F = 3095$ (for descriptors, see Tables 1 and 2). (2) Four models selecting in each case as a test set five compounds: (**3**, **13**, **23**, **33**, **43**), (**5**, **15**, **25**, **35**, **45**), (**6**, **16**, **26**, **36**, **46**) and (**7**, **17**, **27**, **37**, **47**); these data are also given in Table 2. This table contains QSPR statistics for the models with subsequent inclusion into the model of the next best descriptors according to Fischer criterion (up to seven descriptors), on each step the models were checked for the exclusion of previously included descriptors according to this criterion (however in the models given in this paper exclusion of descriptors did not take place). Table 1 shows also the occurrence of the fragmental descriptors for every compound in the model for the whole DS−RI. Figure 1 shows explicitly the change of compound clusters depending on a number of fragmental descriptors.

For DS−NBP[39] we have constructed the following correlations based on the fragmental descriptors: (1) using the whole set of 185 compounds and (2) arbitrarily selecting 19 compounds (**1**,**11**,**21**,**31**,...**171**,**181**), (**3**,**13**,**23**,**33**,...**173**,**183**) and (**5**,**15**,**25**,**35**,...**175**,**185**) for the test set; all these data are given in Table 3.

For DS−BPROH[40] we have constructed QSPRs including from 1 to 6 fragmental descriptors for the same training (70 compounds) and test (30 compounds) sets as in the ref 40. These data are given in Table 4.

For the hydrocarbon data set[41a] we have constructed the QSPR models using both the whole set of 531 compounds (DS-531) (these data are summarized in Table 5) and two different its subsets, namely the subsets of 76 (DS-76) lower alkanes/cycloalkanes and 104 (DS-104) mono- and polycyclic hydrocarbons, which were earlier used for comparative extensive study of structure−boiling point relationships,[21a] the corresponding data are summarized in Table 6.

Fragments as Descriptors in QSPR. Let us first consider the basic methodology of application of fragmental descriptors for QSPR modeling and start from the DS−RI.[38] It consists of 50 substituted phenols with a common scaffold (phenol fragment) and rather small alkyl substituents in different positions of benzene ring. Thus, the structural information is extremely homogeneous.

The authors of ref 38 considered only a training set (supplemented by the leave-one-out cross-validation procedure) and did not use a test set. This methodology is typical for *descriptive* QSPR.[34] For this data set the authors have received a series of "the best" QSPR models using four descriptors, which are different in origin and include, for instance, the combinations of Kier-Hall connectivity and matrix spectrum operators indices. Thus, the interpretation, analysis, and a selection of these QSPR models is difficult, especially taking into account the absence of external proof of the obtained QSPR models. In other words, while the high quality of descriptive stages of QSPR is out of the question, the predictive power of obtained QSPR model/s was not clearly demonstrated. Moreover, taking into account the structural simplicity and similarity of substituents (exclusively alkyl/s) in the whole data set, the following question arises: why do the authors of the paper[38] pay significant attention to weighting schemes for indices applied. Indeed, the account of atom types, which is the real reason for the use of this kind of indices, looks unnecessary for this data set.

In contrast, the selection of fragmental descriptors, which can account for both branch isomerism and positional isomerism, looks quite logical for this case. Even more, one should remember that the property under the study—retention indices—is one of the best for application of various additive schemes (cf., for example, ref 37). Because the additive schemes operate quite successfully for RI, it is another natural reason to select exactly the fragmental descriptors for construction of QSPR models.

Table 2. Statistical Parameters of QSPR Models for Retention Indices of 50 Alkylphenols[38] Based on Fragmental Descriptors[f]

no. of descriptors	1	2	3	4	5	6	7
Training Set: 50 Compounds, No Test Set[a]							
R^2	0.8691	0.9463	0.9825	0.9904	0.9957	0.9976	0.9981
s	44.9	29.1	16.8	12.6	8.5	6.4	5.8
F	319	415	863	1158	2026	2968	3095
max. err.	137.5	72.9	42.3	28.0	27.3	14.2	13.3
Training Set: 45 Compounds; Test Set: Compounds: 3, 13, 23, 33, 43[b]							
R^2	0.8703	0.9458	0.9818	0.9899	0.9956	0.9980	0.9984
s	46	30	18	13	9	6	5
F	288	367	737	980	1781	3158	3378
max. err.	139	73	42	28	28	17	17
mean err.	35	23	13	10	6	4	4
R^2 (test set)	0.8341	0.9461	0.9924	0.9967	0.9958	0.9953	0.9970
max. err. (test set)	65	33	14	8	11	14	9
mean err.(test set)	32	20	7	5	5	4	4
Training Set: 45 Compounds; Test Set: Compounds: 5, 15, 25, 35, 45[c]							
R^2	0.8700	0.9494	0.9834	0.9907	0.9954	0.9974	0.9980
s	45	29	17	13	9	7	6
F	288	394	809	1061	1679	2456	2630
max. err.	137	75	45	27	27	14	13
mean err.	34	21	12	9	6	5	4
R^2 (test set)	0.8553	0.9107	0.9719	0.9869	0.9983	0.9991	0.9982
max. err. (test set)	57	54	27	21	7	4	7
mean err. (test set)	42	30	17	11	4	3	4
Training Set: 45 Compounds; Test Set: Compounds: 6, 16, 26, 36, 46[d]							
R^2	0.8699	0.9502	0.9850	0.9915	0.9957	0.9979	0.9985
s	46	29	16	12	9	6	5
F	288	400	896	1172	1813	3010	3623
max. err.	135	74	42	27	20	15	14
mean err.	37	22	12	9	7	5	4
R^2 (test set)	0.8344	0.8362	0.9061	0.9561	0.9810	0.9855	0.9893
max. err. (test set)	54	36	41	27	17	16	16
mean err. (test set)	23	28	17	12	8	6	5
Training Set: 45 Compounds; Test Set: Compounds: 7, 17, 27, 37, 47[e]							
R^2	0.8694	0.9466	0.9823	0.9904	0.9956	0.9978	0.9982
s	46	30	18	13	9	6	6
F	286	372	760	1036	1783	2810	3008
max. err.	136	73	42	30	27	12	14
mean err.	36	23	13	10	6	5	4
R^2 (test set)	0.8547	0.9380	0.9877	0.9871	0.9961	0.9912	0.9914
max. err. (test set)	44	24	13	15	9	11	15
mean err. (test set)	24	17	6	7	4	6	5

[a] **D1** C_{sp3}; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $C_{Ar}R\div C_{Ar}R$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $CH_2-CH_2-C_{Ar}\div C_{Ar}R$; **D6** $CH_2-C_{Ar}\div C_{Ar}H$; **D7** $CH_2-CH_2-C_{Ar}\div C_{Ar}H$. [b] **D1** C_{sp3}; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $C_{Ar}R\div C_{Ar}R$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $CH_2-CH_2-C_{Ar}\div C_{Ar}R$; **D6** $CH_3-CH_2-C_{Ar}\div C_{Ar}H$; **D7** $C_{sp3}-C_{sp3}-C_{Ar}\div C_{Ar}\div C_{Ar}-C_{sp3}$. [c] **D1** C_{sp3}; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $C_{Ar}R\div C_{Ar}-OH$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $CH_2-CH_2-C_{Ar}\div C_{Ar}R$; **D6** $CH_2-C_{Ar}\div C_{Ar}H$; **D7** $CH_2-CH_2-C_{Ar}\div C_{Ar}H$. [d] **D1** C_{sp3}; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $C_{Ar}R\div C_{Ar}R$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $C_{Ar}H\div C_{Ar}(-CH_2)\div C_{Ar}H$; **D6** $C_{sp3}-C_{sp3}-C_{sp3}-C_{Ar}$; **D7** $C_{sp3}-C_{sp3}-C_{Ar}\div C_{Ar}\div C_{Ar}-C_{sp3}$. [e] **D1** C; **D2** $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$; **D3** $C_{Ar}R\div C_{Ar}R$; **D4** $C_{sp3}-C_{sp3}-C_{Ar}$; **D5** $CH_2-CH_2-C_{Ar}\div C_{Ar}R$; **D6** $CH_2-C_{Ar}\div C_{Ar}H$; **D7** $C_{sp3}-C_{sp3}-C_{sp3}-C_{Ar}$. [f] "C" in the last model means the total number of carbon atoms (both aliphatic and aromatic); "R" means here that aromatic carbon has a substituent.

In fact, the data of Table 2 clearly demonstrate that we have constructed the QSPR models of good quality using only fragmental descriptors. For example, our four-descriptor model for DS−RI has very good statistical parameters ($R^2 = 0.9904$, $s = 13$, $F = 1158$), for the three-descriptor model $R^2 = 0.9825$, $s = 17$, $F = 863$. These data can be compared with data of paper,[38] where the best model with four descriptors has $R^2 = 0.9862$, $s = 15$, $F = 811$ (it should be mentioned however that the authors of paper[38] carried out the orthogonalization and analysis of the descriptors in the model and removed a descriptor having a small contribution, thus a three-descriptor model was obtained with the same correlation coefficient and standard deviation).

Consider the predictive performance of the obtained models. The scientific choice of training vs test set in QSAR/QSPR studies is still an unsolved problem. In principle, the exclusion of some structural information from a training set by extraction of different test sets can lead to different QSAR/QSPR models. We have chosen four test sets (Table 2) by random variation of last digit in numbers of structures in test sets. Inspection of these data leads to the following conclusion: test set statistics (Table 2) demonstrate the stability and predictive power of the fragmental models. Indeed, there are some variations in the sets of best descriptors (selected in each case by a stepwise inclusion of the next best descriptor according to Fischer criterion and exclusion of previously included ones according to this criterion if necessary), but two descriptors are present in all five models, three descriptors in four models, and two descriptors in three models.

Probably the most important reason to apply fragmental descriptors is the "transparency" and interpretability of QSPR models that use structural language. Indeed, the one parameter model uses the descriptor, which is simply the number

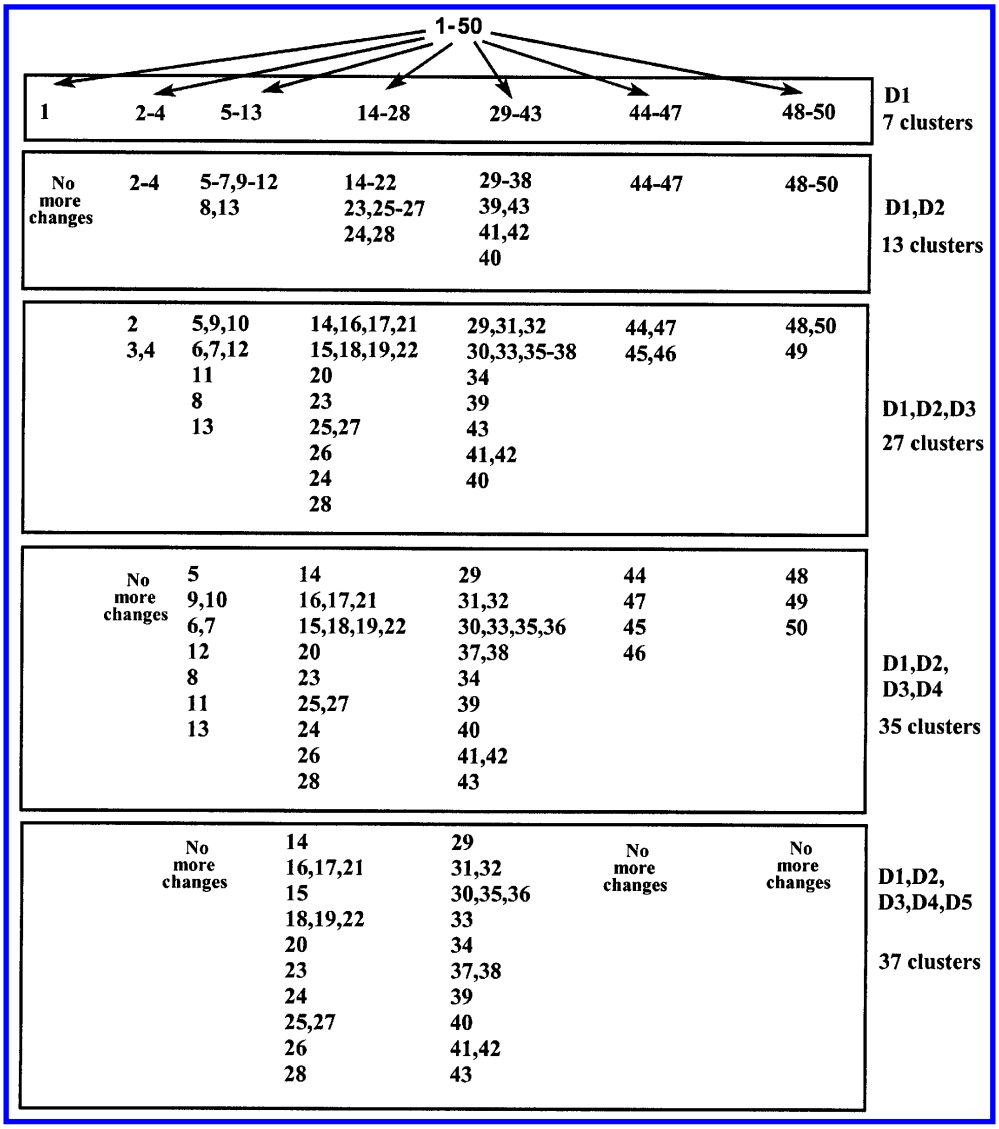

Figure 1. Clustering of compounds from Table 1 by subsequent addition of fragmental descriptors.

of carbon atoms. The plot of experimental (X) vs calculated (Y) values is obviously represented by a set of lines for *isomeric* compounds parallel to *X*-axis. While this primitive index does not provide good statistics (Table 2), Figure 1 demonstrates that this descriptor clusterizes the whole set into seven clusters, containing structural isomers in accordance with the number of carbon atoms in side chain/s. All this is quite natural, because it was well-known that the homology does introduce systematic changes in RI.

The next descriptor, $C_{sp3}-C_{Ar}\div C_{Ar}-C_{sp3}$, transforms the seven clusters into 13 ones (Figure 1), thus sharply improving the statistics (Table 2). Moreover, this clusterization is quite structurally understandable, because it separates, in particular, some *o*-disubstituted compounds, e.g. **8** and **13** vs **5**−**7**, **9**−**12** from the cluster set of **5**−**13** (Figure 1). The third and fourth descriptors again substantially clusterize the set of compounds (up to 35 clusters), giving better and better statistics. On the contrary, the addition of the fifth descriptor gives minor improvement (up to 37 clusters). Analogously, each one of the sixth and seventh descriptors gives one more cluster (**22** vs **18**,**19** and **21** vs **16**,**17,** respectively). This analysis shows that four (or five) descriptors extract the essential structural features of the compounds, and the quality of regression models reaches plateau. The cluster picture shown in Figure 1 nicely clarifies the sensitivity of the used descriptors. It is seen that some clusters still exist even with five descriptors (Tables 1 and 2 and Figure 1), which means that structural difference of isomeric compounds in these clusters cannot be felt by the applied descriptors, and this gives natural restriction for quality of QSPR model/s.[42]

How to treat this important situation? First, if the statistics are good and the difference of experimental values for the compounds in some cluster is small (the best, if it is less than experimental error)—one may ignore the problem.[42] However, one may also (i) add more and more descriptors up to a statistically permitted number or (ii) apply or even create the new ad hoc descriptors to distinguish the compounds in cluster/s. The advantage of fragmental descriptors is that both these operations can be made by a structurally supported way, because one may analyze the structures and account explicitly for the structural difference.

Consider now the QSPR models using fragmental descriptors for DS−NBP[39] (Table 3). It was found before[39] that four parameter models (the molecular connectivity $^1\chi$, *J* modified for the presence of heteroatoms, the electrotopological state *S* of the heteroatoms, and the number N_S of sulfur atoms) give rather good correlation.

Table 3. Statistical Parameters of QSPR Models for Boiling Points of 185 O- and S-Containing Compounds[39] Using Fragmental Descriptors[a]

no. of descriptors	1	2	3	4	5	6	7	8	9	10
Training Set: 185 Compounds, No Test Set[b]										
R^2	0.5244	0.9270	0.9674	0.9733	0.9775	0.9796	0.9814	0.9827	0.9852	0.9876
s	33.3	13.1	8.8	8.0	7.3	7.0	6.7	6.5	6.0	5.5
F	202	1156	1791	1638	1554	1425	1331	1253	1294	1387
max. err.	89.9	44.3	28.7	35.2	33.7	33.2	28.3	25.0	24.2	20.6
Training Set: 166 Compounds; Test Set 1[c]: 19 Compounds (1,11,21,31,41...181)										
R^2	0.4950	0.9248	0.9671	0.9742	0.9783	0.9804	0.9821	0.9838	0.9856	0.9869
s	33.6	13.0	8.6	7.7	7.0	6.7	6.4	6.1	5.8	5.6
F	161	1002	1587	1521	1444	1327	1236	1193	1184	1165
max. err.	87.7	43.9	27.3	26.9	21.5	20.2	18.7	20.7	16.7	17.0
mean err.	28.0	10.2	6.6	5.6	5.2	5.0	4.6	4.4	4.2	4.1
R^2 (test set)	0.6795	0.9365	0.9665	0.9638	0.9694	0.9663	0.9698	0.9727	0.9730	0.9771
max. err. (test set)	65.7	29.2	30.6	38.3	36.6	36.9	36.1	32.0	29.9	27.4
mean err. (test set)	26.7	11.7	6.8	6.8	5.9	6.4	5.8	6.2	6.5	6.0
Training Set: 166 Compounds; Test Set[d]: 19 Compounds (3,13,23,33,43...183)										
R^2	0.5181	0.9272	0.9670	0.9725	0.9778	0.9804	0.9843	0.9862	0.9877	0.9884
s	33.6	13.1	8.8	8.1	7.3	6.9	6.2	5.8	5.5	5.4
F	176	1039	1581	1425	1407	1328	1416	1406	1392	1321
max. err.	90.1	44.1	28.7	35.1	32.2	32.3	26.1	24.4	20.0	17.4
mean err.	28.1	10.1	6.6	5.7	5.2	4.9	4.6	4.2	4.0	3.8
R^2 (test set)	0.5792	0.9246	0.9708	0.9792	0.9714	0.9672	0.9740	0.9715	0.9785	0.9772
max. err. (test set)	63.1	24.1	18.8	12.3	22.5	22.2	17.6	18.5	16.0	16.2
mean err. (test set)	25.0	11.7	7.1	5.5	6.0	6.7	6.4	6.4	5.6	5.8
Training Set: 166 Compounds; Test Set 3[e]: 19 Compounds (5,15,25,35,45...185)										
R^2	0.5233	0.9301	0.9682	0.9747	0.9780	0.9810	0.9831	0.9842	0.9873	0.9881
s	33.0	12.7	8.6	7.7	7.2	6.7	6.3	6.1	5.5	5.4
F	180	1085	1642	1548	1424	1371	1313	1226	1348	1282
max. err.	81.8	44.2	27.2	30.6	28.6	27.6	24.6	23.0	22.9	22.6
mean err.	27.5	9.9	6.7	5.8	5.4	4.8	4.7	4.6	4.0	3.9
R^2 (test set)	0.5308	0.9045	0.9558	0.9634	0.9654	0.9694	0.9688	0.9670	0.9726	0.9709
max. err. (test set)	89.5	31.5	26.8	26.2	26.3	21.3	21.7	21.5	26.1	26.9
mean err. (test set)	30.2	14.0	9.2	7.3	7.5	6.9	6.7	7.1	6.3	6.6

[a] Descriptors in the models (heavy dot means any non-hydrogen atom, R means any substituent except hydrogen atom). [b] **D1** •; **D2** R−S−R; **D3** CH_2; **D4** CH_3−OR; **D5** CH_3−S−CH_2; **D6** CH_3−CH_2−CH_2; **D7** •-•-•-•-•-•; **D8** •-•-•-•-•-•-•-•; **D9** CH_3−CHR_2; **D10** CH_3−SR. [c] **D1** •; **D2** R−S−R; **D3** CH_2; **D4** CH_3−OR; **D5** CH_3−S−CH_2; **D6** CHR_3; **D7** CH_3−CH_2−CH_2; **D8** •-•-•-•-•-•-•-•; **D9** CH_3−CH_2−CHR−OR; **D10** CH_3−SR. [d] **D1** •; **D2** R−S−R; **D3** CH_2; **D4** CH_3−OR; **D5** CH_3−SR; **D6** CH_3−CHR_2; **D7** •-•-•-•-•-•-•-•; **D8** CH_3−CH_2−SR; **D9** •-•-•-•-•-•; **D10** CH_3−O−CH_2−CH_2. [e] **D1** •; **D2** R−S−R; **D3** CH_3; **D4** C_{sp3}−C_{sp3}−C_{sp3}−C_{sp3}; **D5** CH_3−O−CH_2−CH_2; **D6** R−S−S−R; **D7** •-•-•-•-•-•; **D8** •-•-•-•-•-•-•-•; **D9** CH_3−CH_2; **D10** $CH(CH_3)_2OR$.

Table 4. Statistical Parameters of QSPR Models for Boiling Points of the 100 Alcohols Using Fragmental Descriptors[a]

no. of descriptors[b]	1	2	3	4	5	6
R^2	0.8300	0.9602	0.9886	0.9902	0.9917	0.9927
s	12.6	6.1	3.3	3.1	2.9	2.7
F	332	809	1910	1635	1528	1423
max. err.	28.3	16.0	7.1	7.0	7.4	8.0
mean err.	10.2	5.0	2.6	2.4	2.1	1.9
R^2 (test set)	0.7750	0.9523	0.9866	0.9854	0.9852	0.9852
max. err. (test set)	33.0	10.3	9.5	8.9	10.0	10.2
mean err. (test set)	11.1	5.5	2.5	2.6	2.6	2.7

[a] Seventy compounds in a training set and 30 compounds in a test set in accordance with ref 40. [b] Descriptors in the models: **D1** C_{sp3}; **D2** •-•-•; **D3** CH_2−OH; **D4** CH_2−$CH(CH_3)_2$; **D5** CH_3−CHR−OH; **D6** CH_2−CH_2−CH_2−OH.

The data of Table 3 clearly show that a high quality QSPR model can be achieved with four or five fragmental descriptors. While the interpretation of correlation using the set of TIs applied in ref 39 is difficult,[43,44] the involvement of the fragmental descriptors is quite understandable: the combination of the first three descriptors divides the whole set into the clusters with equal numbers of carbon atoms, separates sulfides from other compounds, and accounts for nonbranching carbon atoms. The next descriptors account for the particular features of oxygen- and sulfur-containing structures as well as structural isomerism. In general, the QSPR models (Table 3) constructed using fragmental descriptors exhibit (i) good statistics; (ii) good predictive power as it is demonstrated by the test sets data; and (iii) quite clear explanation and interpretation of the models. We should point out again that the increase in the number of descriptors does not lead automatically to the increase in the quality of a model. As above, the first four-six descriptors have probably picked up the essential structural features.

Consider now the QSPR models using fragmental descriptors for DS−BPROH[40] (Table 4). It was found[40] that one to five parameters (CODESSA; different sets of Randić index (order 1 and 2), informational content of order 1, Kier flexibility index, Kier and Hall index (order 0, 2, 3), and

Table 5. Statistical Parameters of QSPR Models for DS-531 (Boiling Points of 531 Hydrocarbons)[41a]

(a) Topological Indices							
no. of descriptors[a]	1	2	3	4	5	6	7
R^2	0.9596	0.9656	0.9755	0.9784	0.9802	0.9813	0.9825
s	8.4	7.8	6.6	6.2	5.9	5.8	5.6
F	12564	7413	6985	5948	5190	4572	4195
max. err.	61.2	42.9	33.2	34.1	32.4	32.7	31.4
(b) Fragmental Descriptors							
no. of descriptors[b]	1	2	3	4	5	6	7
531 Hydrocarbons (DS-531)							
R^2	0.9069	0.9506	0.9581	0.9604	0.9623	0.9646	0.9662
s	12.8	9.3	8.6	8.4	8.2	7.9	7.7
F	5150	5084	4013	3192	2679	2379	2138
max. err.	90.2	88.3	89.2	89.4	89.4	89.7	86.9
530 Hydrocarbons (Methane Excluded, DS-530)							
R^2	0.9073	0.9553	0.9637	0.9663	0.9684	0.9710	0.9727
s	12.2	8.5	7.6	7.3	7.1	6.8	6.6
F	5169	5625	4648	3766	3208	2914	2655
max. err.	50.4	37.6	40.1	39.5	38.9	38.4	38.4

[a] **D1** $^1\chi$; **D2** W; **D3** $^4\chi_p$; **D4** $^2\chi$; **D5** BIC_1; **D6** J; **D7** $^3\chi_c$. [b] **D1** C_{sp3}; **D2** CH_3; **D3** $C_{sp3}-C_{sp3}(-C_{sp3})-C_{sp3}$; **D4** $CH_3-CHR-CR_2-CH_3$; **D5** $CH_3-CH_2-CR_2-CH_3$; **D6** CH_2-CR_3; **D7** $C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}$ (DS-531) or $CH_3-CH_2-CR_3$ (DS-530).

Table 6. Statistical Parameters of QSPR Models for DS-76 and DS-104 (Boiling Points of Hydrocarbons)[21a]

(a) DS-76 and DS-75 as Training Sets, DS-104 as Test Set (Ref 21a)							
no. of descriptors	1	2	3	4	5	6	7
Training Set of 76 Compounds[a]							
R^2	0.9558	0.9725	0.9767	0.9811	0.9842	0.9858	0.9941
s	12.6	10.1	9.3	8.4	7.8	7.4	4.8
F	1601	1289	1008	923	872	798	1638
max. err.	58.9	58.8	50.6	47.9	43.2	40.6	13.3
mean err.	8.9	6.7	6.4	5.4	5.0	4.6	3.5
R^2 (test set)	0.7466	0.8089	0.1474	0.1367	0.0543	0.1480	0.5515
max. err. (test set)	34.3	30.8	92.1	94.1	94.6	97.7	70.7
mean err. (test set)	11.1	9.0	16.9	17.1	17.7	15.9	11.5
Training Set of 75 Compounds (CH_4 Excluded)[b]							
R^2	0.9624	0.9840	0.9865	0.9884	0.9906	0.9916	0.9935
s	10.3	6.8	6.2	5.8	5.3	5.1	4.5
F	1868	2212	1735	1487	1460	1332	1471
max. err.	29.5	24.4	23.6	20.9	21.0	17.8	11.7
mean err.	7.6	5.3	4.8	4.6	4.0	3.9	3.5
R^2 (test set)	0.7660	0.8253	0.8248	0.8183	0.8305	0.6616	0.6680
max. err. (test set)	33.5	29.9	30.8	31.3	31.2	57.2	64.1
mean err. (test set)	10.5	8.4	8.2	8.5	8.1	10.9	10.4
(b) DS-104 as Training Set (First Five Entries), DS-76 and DS-75 as Test Set (Ref 21a)							
no. of descriptors[c]	1	2	3	4	5	6	7
R^2	0.7955	0.9173	0.9620	0.9680	0.9705	0.9730	0.9758
s	12.2	7.8	5.3	4.9	4.7	4.6	4.3
F	397	560	844	749	644	584	553
max. err.	34.5	29.6	14.5	14.9	13.3	14.2	15.1
mean err.	9.8	5.7	4.0	3.6	3.6	3.4	3.2
R^2 (for test set DS-76)	0.9247	0.9457	0.9483	0.9442	0.9453	0.9452	0.9411
max. err. for DS-76 as test set	93.6	88.3	91.1	95.4	94.3	94.7	98.3
mean err. for DS-76 as test set	10.3	8.9	7.5	7.3	7.2	7.1	7.2
R^2 (for test set DS-75, CH_4 excl.)	0.9444	0.9672	0.9730	0.9715	0.9719	0.9722	0.9703
max. err. for DS-75 as test set (CH_4 excl.)	47.6	30.2	33.1	35.2	33.7	34.0	36.9
mean err. for DS-75 as test set (CH_4 excl.)	9.2	7.8	6.4	6.1	6.1	5.9	6.0

[a] **D1** C_{sp3}; **D2** CH_2; **D3** $C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}$; **D4** CHR_3; **D5** cyclopentyl ring; **D6** $C_{sp3}-C_{sp3}-C_{sp3}$; **D7** CH_3. [b] **D1** C_{sp3}; **D2** CH_2; **D3** $CH_3-CHR-CHR-CH_3$; **D4** $CH_2-CH_2-CH_2-CHR-CH_2-CH_2$; **D5** $CH_3-CR_2-CH_2-CH_2$; **D6** $C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}-C_{sp3}$; **D7** CH_3. [c] **D1** C_{sp3}; **D2** CH_3; **D3** CH_2CR_3; **D4** CH_2CHR_2; **D5** cyclopentyl ring; **D6** cyclopropyl ring; **D7** cyclobutyl ring.

MW[44]) gave rather good correlations. For comparison we present the following statistics obtained:[40] (a) three descriptors (informational content of order 1, Kier flexibility index, Kier and Hall index (order 3)), $R^2 = 0.9754$, RMS = 5.4, $F = 872$ and (b) four descriptors (Randić index (order 1), Randić index (order 2), Kier and Hall index of order 2, Kier and Hall index of order 0), $R^2 = 0.9823$, RMS = 4.91, $F = 899$.

However these data were criticized[40] for (i) lack of interpretation and (ii) not good enough statistics as compared

with the use of variable connectivity index, $^1\chi^f$, which gave $R^2 = 0.982$, RMS = 4.21 (using $x = 0.1$ and $y = -0.92$ as additional adjustable parameters).

The comparison of these data with the data of Table 4 clearly shows that a high quality QSPR model can be achieved with three to five fragmental descriptors. This result is especially good taking into account that application of the variable connectivity index, $^1\chi^f$, means in reality *the hidden* use of two extra parameters in correlation.

Reference 40 contains conceptual criticism of application of usual sets of TIs, while the drawback of application of variable descriptors was not mentioned. However, the interpretation of the regression based on them is not so straightforward.[43] Also, the selection of the parameters *x*, *y*, etc. is still the "try and error" method. What is more, it is not clear, will the optimal parameters be constant if one will change substantially the size of a data set for optimization of these variable parameters. In contrast, the involvement of fragmental descriptors is quite understandable: first descriptor divides the set into clusters of isomers, the second one separates branched and nonbranched chains, and the third one separates primary alcohols, etc.

In general, the QSPR models for DS−BPROH (Table 4) constructed using fragmental descriptors have the same advantages, as it was mentioned above, including good statistics, good predictive power as it is demonstrated by the test sets data, and a quite clear explanation and interpretation of the models. Again, the first three-five descriptors have probably picked up the essential structural features of this structurally simple set.

Consider now the QSPR models based on fragmental descriptors for hydrocarbon data set (DS-531);[41a] these data are shown in Table 5. The ref 41a contains numerous QSPR models for various subsets, ordered by structural features, but we could not find the statistics for the whole data set. Thus, we first performed the comparative constructions of descriptive QSPR models for the whole set DS-531 using (i) the well-known TIs (Table 5a) and (ii) fragmental descriptors (Table 5b). It was done for the following reasons: (a) we just intended to demonstrate that fragmental descriptors are able to give a good *descriptive* model comparable with TI model, and (b) we have recently performed the thorough study of this data set using fragmental descriptors and artificial neural networks and these results will be the subject of special publication. Thus, the QSPR model in this case is just the illustration of the potentials of fragmental descriptors. The data of Table 5 were obtained by stepwise MLR using EMMA package (the pool of indices for selection included more than 200 in the case of TIs and more than 300 in the case of fragments).

First, the comparative analysis of the data of Table 5 (parts a and b) shows that the fragmental descriptors give in general quite acceptable, but definitely a little worse, results in comparison with TIs. One of the reasons for that is evident: the presence of methane in the data set. Indeed, the methane is the outlier for any number of fragmental descriptors (see maximum errors in Table 5b).

In fact, the methane (but the only methane!) represents the specific fragment by itself which cannot be found in any other structure. Thus, we performed the same QSPR modeling but with methane exclusion and these data are given in Table 5b. We should also to comment that this is the usual tactic to improve correlation (see, for example refs 21a and 41b). This exclusion gave visibly better statistics (Table 5b), and the *descriptive* QSPR models either with TIs (Table 5a) or with fragmental descriptors (5b) are quite comparable.

We should also mention here that the plot of experimental vs calculated boiling points in many cases exhibits a noticeable curvature for lower few members, which values are also far from the mean boiling point value of the whole set. In principle, this situation clearly points out a necessity for detailed investigation of nonlinear character of structure-boiling point relationships.[21a,34,41a,45,46] However, we have to emphasize that fragmental descriptors can be easily used in nonlinear equations. Here we only point out the known fact that not the N (N − is the number of carbon atoms) but $N^{2/3}$ has been suggested as the usable variable correlating with boiling points.[46]

Finally we have performed comparative QSPR study for two subsets of hydrocarbons: DS-76 (76 lower alkanes/cycloalkanes) and DS-104 (104 mono- and polycyclic alkanes).[21a] We carried out the QSPR modeling (Table 6), and the predictive performance was estimated as in the ref 21a: the DS-76 was used as training set and the models were checked on the DS-104 and vice versa. Such a selection of training vs test sets is questionable ("rather ill-defined"[21a]), because the structural information in these two sets is very different. At least, definitely structural diversity of the DS-104 is appreciably higher (it contains such structures as spiropentane, bicyclobutane, nortricyclane, and even quadricyclane). However, we have used this procedure simply because it was used before, taking also into account a comparative character of this study.

First, consider the data of Table 6a containing the QSPR models (up to seven descriptors) constructed using DS-76. We were faced here with probably a very typical situation: the *descriptive* QSPR models (up to seven descriptors) are rather good, while, on the other hand, all of them have no *predictive* power for DS-104.

In contrast, the QSPR models (up to seven descriptors) constructed using DS-104 demonstrate good both *descriptive* and *predictive* performances (Table 6b). Indeed, the predictive power, as tested against DS-76, and especially against DS-75 (CH_4 excluded), is fairly good. That means that the structural diversity of DS-104 is enough to construct predictive models for much more simple and homogeneous DS-76; however, the opposite is not true.

One of the reasons for that is understandable: all these models have one extremely pronounced outlier—methane (vide supra). The exclusion of methane and use of DS-75 (75 compounds) for training set visibly improves the statistics (Table 6a): we again obtained fairly good *descriptive* models and *predictive* power of them somewhat improved.

Two points have to be emphasized here. First, as it was mentioned above, a scientific choice of training vs test set in QSAR/QSPR studies is a very important and challenging problem, because the exclusion of some structural information from a training set can lead to a loss of predictive ability for QSAR/QSPR models. Second, it is difficult to expect that the QSPR model, based on the set of simple acyclic and monocyclic structures, can predict properties of strained polycyclic or caged compounds (such as spiropentane, bicyclobutane, nortricyclane, and even quadricyclane present in DS-104). More generally, one should be very careful to

use structurally simple and homogeneous data sets (which is the case in many studies!) for *predictive* QSAR/QSPR modeling of more structurally diverse compounds.

CONCLUSIONS

The fragmental approach, being based on the structural theory, constitutes a universal methodology for predicting various properties of chemical compounds. In distinction with other topological approaches, such as topological indices, structural fragments are easily interpretable and their use is apparent for chemists. Computation of fragmental descriptors does not require the knowledge of the geometry and electronic structure of molecules, and, therefore, they can be used for fast processing of large databases. The success of the fragmental approach is also determined by the diversity of structural fragments as well as by the flexibility of atomic classification. One more advantage of using fragmental descriptors is that selection of them for building QSPR/QSAR models can be made by structurally supported way. What is more, one may expect that the proper classification of stereochemical features incorporated into a fragment structure will permit treatment of stereochemically sensitive properties still being on the topological level.

It should also be pointed out that the use of fragmental descriptors does not exclude application of other types of descriptors. On the contrary, the joint use of fragmental with other types of descriptors often gives better results than their separate application. Actually more than 50 years ago Wiener[47] has shown that the combination of his index with the count of paths of length three gives good results in calculation of boiling points of paraffins.

Finally, it is possible to carry out fast automatic generation of numerous diverse fragmental descriptors, which can further be processed by means of various statistical approaches, such as MLR and PLS and neural networks. The QSAR/QSPR models based on fragmental approach usually include more descriptors than those based on topological indices, but fragmental descriptors are applicable for modeling a wide range of properties and biological activities very often providing good predictive models and interpretable results. In future papers we shall discuss our results of the use of these descriptors with neural networks which makes it possible to apply them in a very efficient way.

The success in applying fragmental descriptors to construction of QSPR/QSAR models relies on effective and flexible algorithm for generating sets of fragments. Such an algorithm underlies the FRAGMENT program, which provides fast generation of numerous types of fragments and supports a flexible hierarchical atomic classification scheme, which will be also the subject of a future publication.

ACKNOWLEDGMENT

The authors thank INTAS (Grant 00-0363) for the support of this work. One of us (N.S.Z) is grateful to the A. v. Humboldt-Foundation for a Research Award and thanks Prof. R. Gleiter for his hospitality at University of Heidelberg.

REFERENCES AND NOTES

(1) (a) Morrison, R. T.; Boyd, R. N. *Organic Chemistry*, 4th ed.; Allyn and Bacon, Inc.: Boston, 1983; Chapter 1.2. (b) We may also provide additional beautiful citation of G. N. Lewis: "No generalization of science, even if we include those capable of exact mathematical statement, has ever achieved a greater success in assembling in simple form a multitude of heterogeneous observations than this group of ideas which we call **structural theory**".

(2) (a) Reinhard, M.; Drefahl, A. *Handbook for Estimating Physicochemical Properties of Organic Compounds*; Wiley: New York, 1999. (b) Wildman, S. A.; Grippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(3) (a) Cox, J. D.; Pilcher, G. *Thermochemistry of Organic and Organometallic Compounds*; Academic Press: London, New York, 1970. (b) Vogel, A. I.; Cresswell, W. T.; Jeffery, G. H.; Leicester, J. Physical Properties and Chemical Constitution. Part XXIV. Aliphatic Aldoximes, Ketoximes, and Ketoxime O−Alkyl Ethers, N, N-Dialkyl Hydrazines, Aliphatic Ketazines, Mono- and Di-alkylaminopropionitriles, Alkoxypropionitriles, Dialkyl Azodiformates, and Dialkyl Carbonates. Bond Parachors, Bond Refractions, and Bond-refraction Coefficients. *J. Chem. Soc.* **1952**, 514−549. (c) Vogel, A. I. *A Textbook of Quantitative Inorganic Analysis*, 3rd ed.; Longmans: London, 1962.

(4) (a) Gibling, T. W. Molecular Volume and Structure. Parts I and II. *J. Chem. Soc.* **1941**, 299−309. (b) Gibling, T. W. Molecular Volume and Structure. Parts III and IV. *J. Chem. Soc.* **1942**, 661−666. (c) Gibling, T. W. Molecular Volume and Structure. Parts V and VI. *J. Chem. Soc.* **1943**, 146−153. (d) Samuel, R. Molecular Constants and Chemical Theories. I. The Parachor (Molecular Volume). *J. Chem. Phys.* **1944**, *12*, 167−179.

(5) (a) Benson, S. W.; Buss, J. H. Additivity Rules for the Estimation of Molecular Properties. Thermodynamic Properties. *J. Chem. Phys.* **1958**, *29*, 546−572. (b) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity Rules for the Estimation of Thermochemical Properties. *Chem. Rev.* **1969**, *69*, 279−324. (c) Bernstein, H. J. The Physical Properties of Molecules in Relation to Their Structure. I. Relations between Additive Molecular Properties in Several Homologous Series. *J. Chem. Phys.* **1952**, *20*, 263−269. (d) Bernstein, H. J. Bond Energies in Hydrocarbons. *Trans. Faraday Soc.* **1962**, *58*, 2285−2306. (e) Franklin, J. L. Prediction of Heat and Free Energies of Organic Compounds. *Ind. Eng. Chem.* **1949**, *41*, 1070−1076. (f) Franklin, J. L. Calculation of the Heats of Formation of Gaseous Free Radicals and Ions. *J. Chem. Phys.* **1952**, *21*, 2029−2033. (g) Sounders, M.; Matthews, C. S.; Hurd, C. O. Relationship of Thermodynamic Properties to Molecular Structure. Heat Capacities and Heat Contents of Hydrocarbon Vapors. *Ind. Eng. Chem.* **1949**, *41*, 1037−1048. (h) Sounders, M.; Matthews, C. S.; Hurd, C. O. Entropy and Heat of Formation of Hydrocarbon Vapors. *Ind. Eng. Chem.* **1949**, *41*, 1048−1056. (i) Zahn, C. T. The Significance of Chemical Bond Energies. *J. Chem. Phys.* **1934**, *2*, 671−680.

(6) (a) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180. (b) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficient. *Nature* **1962**, *194*, 178−180. (c) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(7) (a) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationship. *Acc. Chem. Res.* **1969**, *2*, 232−239. (b) Hansch, C.; Leo, A. Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. *ACS Professional Ref. Book* **1995**.

(8) (a) Nys, G. G.; Rekker, R. F. Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. Introduction of Hydrophobic Fragmental Constants (f-Values). *Eur. J. Med. Chem. − Chim. Therap.* **1973**, *8*, 521−535. (b) Nys, G. G.; Rekker, R. F. The Concept of Hydrophobic Fragmental Constants (f-Values). II. Extension of its Applicability to the Calculation of Lipophilicities of Aromatic and Heteroaromatic Structures. *Eur. J. Med. Chem. − Chim. Therap.* **1974**, *9*, 361−375. (c) Rekker, R. F.; de Kort, H. M. The Hydrophobic Fragmental Constants; an Extension to a 1000 Data Point Set. *Eur. J. Med. Chem.* **1979**, *14*, 479−488.

(9) (a) Tatevsky, V. M. *The Chemical Structure of Hydrocarbons and Regularities in their Physico-Chemical Properties (Rus)*; Moscow University: Moscow, 1953. (b) Tatevsky, V. M.; Bendersky, V. A.; Yarovoi, S. S. *The Chemical Structure of Hydrocarbons and Regularities in their Physico-Chemical Properties (Rus)*; Moscow University: Moscow, 1960. (b) Tatevsky, V. M. *The Classical Theory of Molecular Structure and Quantum Mechanics (Rus)*. Khimiya: Moscow, 1973. (c) Tatevsky, V. M. *The Theory of Physico-Chemical Properties of Molecules and Substances (Rus)*; Moscow University: Moscow, 1987.

(10) Kubinyi, H. The Quantitative Analysis of Structure−Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 5th ed.; Wolff, M. E., Ed.; Wiley: New York, 1995; Vol. 1, pp 497−571.
(11) Smolensky, E. A. Application of the Graph Theory to Calculation of Structurally-Additive Properties of Hydrocarbons. *Zh. Fiz. Khim.* **1964**, *38*, 1288−1291.
(12) (a) Smolensky, E. A. A Semiempirical Calculation Methods of Formation Energies of Alkanes. *Dokl. Akad. Nauk SSSR (Russ.)* **1976**, *230*, 373−376. (b) Smolensky, E. A.; Kocharova, L. V. The Thermochemical Conformational Analysis and Calculation of Standard Formation Enthalpies of Cl−Substituted Alkanes. *Dokl. Akad. Nauk SSSR (Russ.)* **1982**, *261*, 112−115.
(13) Of course, the model can be more complex and sophisticated, including, say, quadratic or nonadditive terms.
(14) Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. On the Basis of Invariants of Labeled Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 527−531.
(15) (a) Cammarata, A. Interrelationship of the Regression Models Used for Structure-Activity Analyses. *J. Med. Chem.* **1972**, *15*, 573−577. (b) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure−Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399. (b) Fujita, T.; Ban, T. Structure−Activity Study of Phenethylamines as Substrates of Biosynthetic Enzymes of Sympathetic Transmitters. *J. Med. Chem.* **1971**, *14*, 148−152.
(16) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, Germany, 1993.
(17) (a) Gautzsch, R.; Zinn, P. List Operations on Chemical Graphs. 5. Implementation of Breadth-First Molecular Path Generation and Application in the Estimation of Retention Index Data and Boiling Points. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 791−800. (b) Platt, J. R. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* **1947**, *15*, 419. (c) Randić, M. Graph Valence Shells as Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 627−630.
(18) Randić, M.; Zupan, J. On interpretation of Well-Known Topological Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550−560.
(19) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750. (b) Viswanadhan, V. N.; Ghose, A. K.; Singh, U. C.; Wendoloski, J. J. Prediction of Solvation Free Energies of Small Organic Molecules: Additive-Constitutive Models Based on Molecular Fingerprints and Atomic Constants. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 405−412. (c) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746−753.
(20) Rücker, G.; Rücker, C. Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683−695. (b) Diudea, M. V.; Minailiuc, O. M.; Katona, G. Novel Connectivity Descriptors Based on Walk Degrees. *Croat. Chem. Acta* **1996**, *69*, 857−871. (c) Gutman, I.; Rücker, C.; Rücker, G. On Walks in Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 739−745.
(21) (a) Lučić, B.; Lukovits, I.; Nikolić, S.; Trinajstić, N. J. Distance-Related Indexes in the Quantitative Structure − Property Relationship Modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 527−535. (b) Rücker, G.; Rücker, C. Symmetry-Aided Computation of the Detour Matrix in Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 710−714.
(22) (a) Avidon, V. V. The Comparison Criteria of Chemical Structures and the Principles of Designing a Description Language for Processing Chemical Information on Biologically Active Compounds. *Khim. −Farm. Zhurnal (Russ.)* **1974**, *8*, 22−25. (b) Avidon, V. V.; Arolovich, V. S.; Kozlova, S. P. A Statistical Study of the Information File on Biologically Active Compounds. II. Prediction of Biological Activity Through Substructural Analysis. *Khim. −Farm. Zhurnal (Russ.)* **1978**, *12*, 88−92. (c) Avidon, V. V.; Pomerantsev, I. A.; Rozenblit, A. B.; Golender, V. E. Structure−Activity Relationships Oriented Languages for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 207−214.
(23) (a) Filimonov, D. A.; Poroikov, V. V.; Karaicheva, E. I. Computer-Aided Prediction of Prodrug Activity Spectra for Chemical Substances on the Basis of their Structural Formulas: Computerized System PASS. *Eksperiment. Klin. Farm. (Russ.).* **1995**, *58*, 56−62. (b) Franke, R.; Huebel, S.; Streich, W. J. Substructural QSAR Approaches and Topological Pharmacophores. *Environ. Health. Perspect.* **1985**, *61*, 239−255. (c) Kotovskaya, S. K.; Tyurina, L. A.; Chernova, E. Y.; Mokrushina, G. A.; Chupakhin, O. N.; Novikova, A. P.; Ilyenko, V. I. Algorithmic Search for Compounds with Antiviral Activity in a Set of Nitrogen- and Sulfur-Containing Heterocycles. *Khim. −Farm. Zhurnal (Russ.)* **1989**, *22*, 310−314. (d) Rozenblit, A. B.; Golender, V. E. *Logical Combinatorial Algorithms for Drug Design*; Research Studies Press, Wiley & Sons: New York, Chichester, Brisbane, Toronto, 1983. (e) Poroikov, V. V.; Filimonov, D. A.; Borodina, Yu. V.; Lagunin, A. A.; Kos, A. Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355. (f) Filimonov, D. A.; Poroikov, V. V.; Borodina, Yu. V.; Gloriosova, T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.
(24) (a) Cunningham, A. R.; Klopman, G.; Rosenkranz, H. S. Identification of Structural Features and Associated Mechanisms of Action for Carcinogens in Rats. *Mutat. Res.* **1998**, *405*, 9−27. (b) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321. (c) Klopman, G.; Namboordiri, K.; Schochet, M. Simple Method of Computing the Partition Coefficient. *J. Comput. Chem.* **1985**, *6*, 28−38. (d) Klopman, G.; Wang, S. A Computer Automated Structure Evaluation (CASE). Approach to Calculation of Partition Coefficient. *J. Comput. Chem.* **1991**, *12*, 1025−1032. (e) Klopman, G.; Rosenkranz, H. S. Prediction of Carcinogenicity/ Mutagenicity Using MultiCASE. *Mutat. Res.* **1994**, *305*, 33−46. (f) Klopman, G.; Li, J.; Wang, S.; Dimayuga, M. Computer Automated log P Calculations Based on an Extended Group Approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(25) Meylan, W. M.; Howard, P. H. Atom/Fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(26) (a) Brinn, M.; Payne, M. P.; Walsh, P. T. Neural Network Prediction of Mutagenicity Using Structure−Property Relationships. *Chem. Eng. Res. Des.* **1993**, *71*, 337−339. (b) Brinn, M.; Walsh, P. T.; Payne, M. P.; Bott, B. Neural Network Classification of Mutagens Using Structural Fragment Data. *SAR QSAR Environ. Res.* **1993**, *1*, 169−210.
(27) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Computational Neural Networks as an Alternative to the Linear Regression Analysis in the Studies of Quantitative Structure−Property Relationships for the Case of Physico-Chemical Properties of Hydrocarbons. *Dokl. Akad. Nauk (Russ.)* **1993**, *332*, 713−716.
(28) Burden, F. R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct. −Act. Relat.* **1996**, *15*, 7−11.
(29) (a) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Godette, T. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of Quantitative Structure−Activity Relationship Methods for Large-scale Prediction of Chemical Binding to the Estrogen Receptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677. (b) Burden, F. R.; Winkler, D. A. New QSAR Methods Applied to Structure Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 236−242.
(30) (a) Kumskov, M. I.; Ponomareva, L. A.; Smolensky, E. A.; Mityushev, D. F.; Zefirov, N. S. Method of automatic generation of structural descriptors of organic compounds for QSAR. *Izv. Akad. Nauk (Russ.)* **1994**, 1391−1393. (b) Kumskov, M. I.; Mityushev, D. F. Group Method of Data Handling (GMDH) as Applied to Collective Property Estimation of Organic Compounds by an Inductive Search of Their Structural Spectra. *Pattern Recognition Image Analysis* **1996**, *6*, 497−509. (c) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of Ion Complexation and Extraction Using Substructural Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847−858.
(31) (a) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Computer Program for Generating Sets of Subgraphs for Molecular Graphs. *Abstracts of Papers of the Conference "Molecular Graphs in Chemical Studies"*; Kalinin, 1990; p 5. (b) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. A Method for Automatic Selection of Structural Fragments in the Search for Structure−Property Correlations Based on their Hierarchic Classification. *Proceedings of the 1st All-Union Conference on Theoretical Organic Chemistry*; Volgograd, 1991; p 557. (c) Palyulin, V. A.; Baskin, I. I.; Petelin, D. E.; Zefirov, N. S. Novel Descriptors of Molecular Structure in QSAR and QSPR Studies. *Abstracts of the 10th European Symposium on Structure−Activity Relationships: QSAR and Molecular Modelling*; Barcelona, 1994; B257. (d) Palyulin, V. A.; Baskin, I. I.; Petelin, D. E.; Zefirov, N. S. Novel Descriptors of Molecular Structure in QSAR and QSPR Studies. *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona, 1995,; pp 51−52. (e) Palyulin, V. A.; Radchenko, E. V.; Baskin, I. I.; Zotov, A. Yu.; Zefirov, N. S. Substructural and Superstructural Approaches in QSAR. *Abstracts of the 11th European Symposium on QSAR: Computer assisted lead finding and optimization*; Lausanne, 1996; p 31.A.
(32) (a) Petelin, D. E.; Palyulin, V. A.; Zefirov, N. S. Topological Indexes Based on Weights of the Molecular Graph Vertexes for QSAR and QSPR Studies. *Dokl. Akad. Nauk (Russ.)* **1992**, *324*, 1019−1022 (*Chem. Abstr.*, *118*, 38223). (b) Pivina, T. S.; Sukhachev, D. V.;

Maslova, L. K.; Shlyapochnikov, V. A.; Zefirov, N. S. Studies of Correlations for Structure-Thermal Stability Parameters of Nitro Compounds Based on the QSPR Approach. *Dokl. Akad. Nauk (Russ.)* **1993**, *330*, 468−472 (*Chem. Abstr*., *120*, 54070). (c) Sukhachev, D. V.; Pivina, T. S.; Shlyapochnikov, V. A.; Petrov, E. A.; Palyulin, V. A.; Zefirov, N. S. Study of Quantitative Structure-Drop-Weight-Impact-Sensitivity Relationships for Organic Polynitro Compounds. *Dokl. Akad. Nauk* (*Russ.)* **1993**, *328*, 188−189 (*Chem. Abstr*., *118*, 257697). (d) Sukhachev, D. V.; Pivina, T. S.; Zhokhova, N. I.; Zefirov, N. S.; Zeman, S. QSPR Approach to Calculating the Rate Constants of Homolysis of Nitro Compounds in Different Aggregation States. 1. Gas State. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1995**, 1653−1656 (*Chem. Abstr*., *124*, 260218). (e) Sukhachev, D. V.; Pivina, T. S.; Zhokhova, N. I.; Zefirov, N. S.; Zeman, S. QSPR Approach for Calculating the Rate Constants of Homolysis of Nitro Compounds in Different Aggregation States. 2. Liquid State. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1995**, 1657−1660 (*Chem. Abstr*., *124*, 260219). (f) Sukhachev, D. V.; Pivina, T. S.; Zhokhova, N. I.; Zefirov, N. S.; Zeman, S. QSPR Approach for Calculating the Rate Constants of Homolysis of Nitro Compounds in Different Aggregation States. 3. Solid State. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1995**, 1661−1665 (*Chem. Abstr*., *124*, 260220).

(33) Zefirov, N. S.; Petelin, D. E.; Palyulin, V. A.; McFarland, J. W. Quantitative relationship between the structure of 2-substituted 1,2,4-triazine-3, 5(2H, 4H)-diones and their anticoccidial activity. *Dokl. Akad. Nauk (Russ.)* **1992**, *327*, 504−508.

(34) Zefirov, N. S.; Palyulin, V. A. QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models? *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1022−1027.

(35) (a) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. NASA. A Computer Program for Performing QSAR/QSPR Studies Using Artificial Neural Networks. In *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications*; Sanz, F., Giraldo, J., Manaut, F., Eds.; Prous Science Publishers: Barcelona, 1995; pp 30−31. (b) Halberstam, N. M.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. NASAWIN − A Program Simulator of Neural Networks for Structure−Activity Relationship Studies. In *International symposium CACR-96;* Book of Abstracts, Moscow, 1996; pp 37−38.

(36) Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. A Neural Device for Searching Direct Correlations between Structures and Properties of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 715−721.

(37) (a) Golovnya, R. V.; Garbuzov, V. G.; Misharina, T. A. Gas-chromatografic Characteristic of Sulfur Containing Compounds. 3. *n*-Alkyl-iso-alkyl Sulfides (Russ.). *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1978**, 387−397. (b) Golovnya, R. V.; Misharina, T. A.; Garbuzov, V. G. Gas-chromatografic Characteristic of Sulfur Containing Compounds. 6. Di-iso-Aliphatic sulfides (Russ.). *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1979**, 1029−1032. (c) Martinu, V.; Janak, J. Gas−Liquid Chromatographic Retention Data of Some Aliphatic and Alicyclic Sulphides. *J. Chromatogr*. **1970**, *52*, 69−75. (d) Morishita, F.; Murakita, H.; Takemura, Y.; Kojima, T. Prediction of Molecular Structures of Thiols and Sulphides by Retention Indices. *J. Chromatogr*. **1982**, *239*, 483−492.

(38) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Comparison of Weighting Schemes for Molecular Graph Descriptors: Application in Quantitative Structure − Retention Relationship Models for Alkylphenols in Gas−Liquid Chromatography. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732−743.

(39) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlation between Chemical Structure and Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals, and Their Sulfur Analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237−244.

(40) Randić, M.; Pompe, M. The Variable Connectivity Index $^1\chi^f$ versus the Traditional Molecular Descriptors: A Comparative Study of $^1\chi^f$ Against Descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 631−638.

(41) (a) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 788−802. (b) The referee comment about accounting of methane is of interest: "Exclusion of methane is not a dirty trick to obtain optically better correlations, often it is justified by good reason, since inclusion of methane would include information on a compound far outside the size range one is really interested. Sometimes exclusion of methane even seems to worsen a correlation., e.g. in cases where a high R^2 value is obtained simply because the compounds treated span a large size range, and so the TIs used will span a wide numerical range. Further, in cases where logarithms of TIs used, methane with its notoriously numerous TI values equal to zero necessarily has to be excluded".

(42) We must point out here another ignored problem, which put natural restriction for all topological indices: it is the presence of diastereomers, having different boiling points but the same value of indices.

(43) Interpretation of well-known topological indices (ref 18) is based on "... structural concepts, which are well understood and need no explanation". However, evidently this statement entirely fits for fragmental descriptors!

(44) The appearance of either the number N_S of sulfur atoms (ref 39) or MW (ref 40) as the descriptors is remarkable!

(45) Randić, M.; Pompe, M. The Variable Molecular Descriptors Based on Distance Matrices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 575−581.

(46) Cao, C.; Liu, S.; Li, Z. On Molecular Polarizability: 2. Relationship to the Boiling Point of Alkanes and Alcohols. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1105−1111.

(47) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

CI020010E