QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models?

N. S. Zefirov* and V. A. Palyulin

Department of Chemistry, Moscow State University, Moscow, 119899 Russia

Received January 3, 2001

Our investigation was motivated by a recent paper concerning "high-quality structure-property-activity regressions" (*J. Chem. Inf. Comput. Sci*. **2000**, *40*, 899−905) for boiling points of small sulfides using variable connectivity indices. We performed QSAR study of the same data set using a logically preselected solvation index and obtained a very good one-parameter regression. The structures of the whole possible set of small sulfides (C2−C6) were generated and the statistics were proven by real ***prediction*** using an external test set of sulfides. The variants of extended prediction with extrapolated data and QSAR using an expanded training set were also performed, and all these data also revealed the preference of the solvation index. A general problem of *descriptive* vs *predictive* QSAR is discussed.

INTRODUCTION

At first, QSAR was the attempt to express relative biological response in terms of a set of physical variables, representing hydrophobic, electronic, and steric factors.[1] In the broadest sense, a QSAR may be nowadays regarded as a computer derived rule/s which quantitatively describe/s any property (e.g. biological activity) in terms of descriptors of chemical structures.[1−3]

The QSAR study includes the following principal steps:

1. Selection of a "training set" (the set of compounds for construction of correlation equation/s) and a "test set" (the set to prove the predictive power of correlation/s) from the whole available set of compounds with known activity/property.

2. Selection or choice of a set of descriptors, which are believed to adequately characterize the structures and are somehow relevant to a property to be estimated.

3. Tracing the correlation of the property under investigation vs descriptors using statistical methods, creation, and validation of "QSAR model/s".

4. The proof of the constructed model/s using the "test set" to select the model/s having predictive power.

5. Use of the obtained QSAR model for a prediction study.

The results of steps 2 and 3 are *descriptive* QSAR models. In a sense, the steps 1−3 are only preliminary steps if to consider as a target the real search for active compounds. Step 4 gives *predictive* QSAR model/s, which have to be used for step 5. This step includes either (a) a generation of novel structures with subsequent selection of those with the desired values of properties using a QSAR model as a filter and/or (b) a generation of novel structures with a given value of property using a correlating equation ("inverse QSAR"[4−8]).

Nowadays, the QSAR approach receives enormously wide applications. It is often used routinely for predicting numerous physicochemical properties and biological activity of organic compounds and for the computer-assisted design of novel structures with given properties.

Unfortunately, many methodological problems of QSAR are still poorly understood or even remain unsolved, including such important ones as (i) scientific choice of "training" vs "test" set of compounds; (ii) selection of proper descriptors from the great pool of different indices; (iii) selection of "the best" QSAR model; and (iv) applicability of QSAR to nonhomogenious sets of compounds. In fact, a majority of graph theoretical constructions were proven only for selected properties of simple hydrocarbons (even for them mp is exclusion![9]) or for their rather simple heteroatomic derivatives,[10] etc. It means, that for many occasions the QSAR should be considered as a sort of art; very often the strictly made calculations are mixed with the intuition-based determination of the applicability range of QSAR equations or arbitrary selection of descriptors. The predictions based on constructed QSAR models are very rarely published (if made at all); even the stage of model validation using a test set is often omitted.

The purpose of this paper concerns the problems (a) of a selection of high-quality QSAR models and (b) proper selection of a descriptor for the particular set of data. This study was motivated by recent paper of Randić and Basak[12] because these authors especially address themselves to "... the problem of construction of ***high-quality*** regressions (HQR)...". Moreover, because this paper is written by leading scientists in the field of topological indices, the comparative analysis should give especially valuable methodological information.

First, we have to emphasize that the methodology used in the paper[12] is typical for descriptive QSAR: the authors[12] consider only a training set and do not use a test set. Hence, all conclusions about the "high quality of the QSAR correlations" were extracted exclusively from statistics criteria (mainly, standard errors), and this paper has no any *external* proof of the obtained QSAR models. In other words, while the high quality of *descriptive* stages of QSAR is out of the question, the *predictive* power of obtained QSAR model/s was not demonstrated.

Second, the paper[12] advocates the use of one particular type of topological indices, namely of "variable molecular

* Corresponding author e-mail: zefirov@org.chem.msu.ru and vap@org.chem.msu.su.

10.1021/ci0001637 CCC: $20.00
Published on Web 05/12/2001

Table 1. Alkyl Sulfides Used in Present Work and Their Boiling Points[a] and Index Values

				connectivity indices					
no.	no. of carbons	sulfide	bp, C°	$^1\chi$	$^1\chi^s$	$^1\chi^f$	Z	S_w	W
1	C-2	dimethyl sulfide	37.3	1.41421	2.12132	1.74574	3	3.186	4
2	C-3	methyl ethyl sulfide	66.6	1.91421	2.51777	2.11976	5	4.493	10
3	C-4	methyl propyl sulfide	95.5	2.41421	3.01777	2.56420	8	5.888	20
4	C-4	diethyl sulfide	92.0	2.41421	2.91421	2.49377	8	5.888	20
5	C-4	methyl isopropyl sulfide	84.4	2.27006	2.82773	2.40648	7	5.906	18
6	C-5	ethyl isopropyl sulfide	107.4	2.77006	3.22418	2.78049	11	7.511	32
7	C-5	methyl butyl sulfide	123.2	2.91421	3.51777	3.00865	13	7.351	35
8	C-5	methyl isobutyl sulfide	112.5	2.77006	3.37361	2.88555	11	7.351	32
9	C-5	ethyl propyl sulfide	118.5	2.91421	3.41421	2.93821	13	7.351	35
10	C-5	methyl *tert*-butyl sulfide	101.5	2.56066	3.09099	2.64784	9	7.366	28
11	C-6	methyl amyl sulfide	145.0	3.41421	4.01777	3.45309	21	8.869	56
12	C-6	ethyl butyl sulfide	144.2	3.41421	3.91421	3.38266	21	8.869	56
13	C-6	dipropyl sulfide	142.8	3.41421	3.91421	3.38266	21	8.869	56
14	C-6	propyl isopropyl sulfide	132.0	3.27006	3.72418	3.22494	18	9.130	52
15	C-6	ethyl isobutyl sulfide	134.2	3.27006	3.77006	3.25956	18	9.130	52
16	C-6	methyl isoamyl sulfide	137.0	3.27006	3.87361	3.32999	18	9.130	52
17	C-6	methyl 2-methylbutyl sulfide	139.0	3.30806	3.91161	3.35550	19	8.999	50
18	C-6	ethyl s-butyl sulfide	133.6	3.30806	3.76218	3.25044	19	8.999	50
19	C-6	ethyl *tert*-butyl sulfide	120.4	3.06066	3.48744	3.02185	14	9.278	46
20	C-6	diisopropyl sulfide	120.0	3.12590	3.53415	3.06722	15	8.052	48
21	C-6	methyl 1-ethylpropyl sulfide	137.0	3.34607	3.90374	3.34637	20	8.498	48
22	C-5	methyl s-butyl sulfide (test)[b]	114.5	2.80806	3.36574	2.87643	12	7.238	31
23	C-6	methyl *tert*-amyl sulfide (test)[b]	128.3	3.12132	3.65165	3.13364	16	8.644	44
24	C-6	methyl s-amyl sulfide[b]		3.30806	3.86574	3.32087	19	8.999	50
25	C-6	methyl 1,2-dimethylpropyl sulfide (test)[b]	133.0	3.18074	3.73842	3.21031	17	8.764	46
26	C-6	methyl 2,2-dimethylpropyl sulfide[b]		3.06066	3.66421	3.14844	14	9.277	46
27	C-7	methyl hexyl sulfide	171.0	3.91421	4.51777	3.89753	34	10.433	84
28	C-7	propyl butyl sulfide	166.0	3.91421	4.41421	3.82710	34	10.433	84
29	C-7	propyl isobutyl sulfide	155.0	3.77006	4.27005	3.70401	29	10.766	79
30	C-7	isopropyl isobutyl sulfide	145.0	3.62590	4.08002	3.54629	25	9.403	74
31	C-7	ethyl 2-methylbutyl sulfide	159.0	3.80806	4.30806	3.72951	31	10.748	76
32	C-7	propyl *tert*-butyl sulfide	138.0	3.56066	3.98744	3.46629	23	11.142	71
33	C-7	isopropyl s-butyl sulfide	142.0	3.66390	4.07215	3.53717	26	10.161	71
34	C-7	ethyl isoamyl sulfide	159.0	3.77006	4.27005	3.70401	29	10.766	79
35	C-7	isopropyl butyl sulfide	163.5	3.77006	4.22418	3.66938	29	10.766	79
36	C-8	dibutyl sulfide	188.9	4.41421	4.91421	4.27155	55	12.036	120
37	C-8	diisobutyl sulfide	170.0	4.12590	4.62590	4.02536	40	10.777	108
38	C-8	butyl isobutyl sulfide	178.0	4.27005	4.77005	4.14845	47	12.422	114
39	C-8	di-*tert*-butyl sulfide	148.5	3.70711	4.06066	3.54993	24	10.166	88
40	C-8	di-s-butyl sulfide	165.0	4.20191	4.61016	4.00711	45	11.616	100
41	C-8	butyl s-butyl sulfide	177.0	4.30806	4.76218	4.13933	50	12.492	110
42	C-8	isobutyl s-butyl sulfide	167.0	4.16390	4.61803	4.01623	43	11.570	104
43	C-8	methyl heptyl sulfide	195.0	4.41421	5.01777	4.34198	55	12.036	120
44	C-7	isopropyl *tert*-butyl sulfide (test)	128.5	3.41650	3.79740	3.30857	19	9.644	66
45	C-8	isopropyl *tert*-amyl sulfide (test)	160.5	3.97716	4.35806	3.79438	34	11.320	92

[a] **1**−**21** and **27**−**43** from ref 13; **22**, **23**, **25** from ref 37; **44** and **45** from ref 40. [b] The structures generated with SMOG.[35,36]

descriptors" (in reality, the variable connectivity index $^1\chi(x,y)$), arguing that it has been the index which produces appreciably better statistics. It is important, because a general issue of QSAR is the selection of adequate descriptors. However, the following question may be raised: to what extent the selection of variable connectivity index in this case is (i) logically justified and (ii) really proven by the obtained results.

Thus, we will touch on the important methodological problems of QSAR, even focusing on the above-mentioned particular questions.

RESULTS AND DISCUSSION

Following the paper[12] we used the data set which consists of 21 dialkyl sulfides (actual data were taken from ref 13), containing from two to six carbon atoms; they are shown in Table 1 (compounds **1**−**21**[14]). The considered structural data are extremely homogeneous, because they include only acyclic structures containing exactly one heteroatom—sulfur. Thus, the authors of the paper[12] really faced only with the problem to differentiate the atom types, namely carbon and sulfur [for instance, to distinguish methyl amyl sulfide (**11**) vs ethyl butyl sulfide (**12**)]. The solution was thought to be found in the application of variable connectivity indices.

However, the obvious question can be raised: as far as the many types of weighted topological indices are known which may provide this type of differentiation, why does one need to prefer exactly the variable index. Indeed, one should clearly understand that the variable index $^1\chi^f$, recommended for the considered case, contains two extra variables (x for C and y for S[12]). Hence, the QSAR correlation obtained is, in fact, a *hidden three-parametric* model. In other words, the problem is formulated as the following: is it possible to select some of the weighted indices, which may provide a real *one-parameter* correlation.

Table 2. QSAR Models (bp = $ax + b$) for Boiling Points of 21 Simple Sulfides (Compounds **1**−**21**) and Their Statistical Parameters

index	W	S_w	Z	$^1\chi$	$^1\chi^s$	$^1\chi^f$
a	1.652	15.77	4.676	50.75	54.05	60.20
b	52.50	−4.930	50.42	−30.92	−69.46	−61.33
R^2	0.9210	0.9145	0.9063	0.9868	0.9887	0.9918
s	8.1	8.4	8.8	3.3	3.1	2.6
F	222	203	184	1416	1664	2291
max. error	21.8	21.0	27.1	7.7	7.9	6.5
Test Set: Compounds **22, 23, 25**						
RMS	7.0	3.5	5.2	2.3	1.2	1.8
max. pred. err	10.8	5.3	8.0	2.9	2.0	2.7
Test Set: Compounds **22, 23, 25, 27**−**45**						
RMS	40.1	11.7	60.0	8.5	7.4	7.7
max. pred. err	62.8	32.8	118.7	17.3	14.7	14.9

We have examined these problems using for correlation our so-called "solvation index", $^1\chi^s$, which we introduced in 1991 to treat the enthalpies of nonspecific solvation.[15,16] Indeed, the solvation enthalpy of, say, propane, CH_3CH_2-CH_3, and dimethylmercury, CH_3HgCH_3, differs enormously, but both of these molecules are represented by the same hydrogen depleted graph, and, hence, have the identical topological indices which do not take into account atom types. Our solvation index was created exactly to differentiate such cases.

The general formula for the solvation index (calculated for hydrogen- and fluorine-depleted molecular graphs) is the following

$$^m\chi^s = (1/2^{m+1}) \Sigma Z_i Z_j ... Z_k / (\delta_i \delta_j ... \delta_k)^{1/2}$$

where m is the order of index; summation is over all subgraphs of order m; δ_i, δ_j, ... δ_k are connectivities of vertexes of subgraph; and Z_i, Z_j, ... Z_k are coefficients characterizing the atom size, which we accepted to be equal to the number of the period in the Periodic Table. The term $1/2^{m+1}$ just normalizes values of $^m\chi^s$ to provide their coincidence with $^m\chi$ for the elements of the second row. In this study we have used the simplest solvation index of the first order

$$^1\chi^s = 0.25 \Sigma Z_i Z_j / (\delta_i \delta_j)^{1/2}$$

with $Z = 2$ for carbon and $Z = 3$ for sulfur (number of a corresponding period in the Periodic Table) with summation over all bonds.

We have to mention that there exist more detailed weighting schemes which take into account atomic number, electronegativity, relative covalent radius, etc.[17−20] However, our index is probably the simplest one and differentiates C and S in the most economic way, which perfectly corresponds to the problem under investigation.

We carried out a QSAR study of a data set of 21 sulfides using single topological descriptors (see values in Table 1), selecting those which can give a reasonable correlation.[21] Among them we have found the Hosoya[9,29] (Z), Wiener[30,31] (W), and Randić[32] ($^1\chi$) indices[33] (Table 2). We additionally reproduced the data[12] using a variable index $^1\chi^f$, $x = 0.25$, and $y = -0.95$; the result is also presented in Table 2.[34] We included also the correlation data using one of "vertex-weights" (S_w) indices[27] (see Appendix), which are comparable with data for Wiener index (Table 2).

The results of Table 2 show that several topological indices alone reproduce the boiling point values reasonably well. That means, in fact, that almost all of these indices are very intercorrelated. However W, Z, and S_w indices give worse statistics, than connectivity indices. Inspection of Table 2 clearly demonstrates that our solvation index $^1\chi^s$ is really good for correlation purposes. If we accept the standard deviation as a preferential criterion of model quality in accordance with ref 12, then the solvation index is better than $^1\chi$ and quite comparable to $^1\chi^f$ (However, we should remember again that the latter has, in fact, two more adjustable parameters!).

To be "good", any graph theoretical index should at least have good discrimination for isomeric compounds; otherwise, the preciseness of regressions cannot be better than the property differences of isomeric compounds.[11] Inspection of Table 1 reveals that, as expected, $^1\chi$ (as well as W, S_w, and Z) is degenerative in many cases, and, hence, the precision of regressions based on them is limited. On the other hand, the sensitivity of both $^1\chi^s$ and $^1\chi^f$ is high but not thorough, and degeneracy occurs only for two compounds **12** and **13**. This indicates approximately equal sensitivity or discriminative power of both indices.

Thus, the data obtained show that (a) statistics differences are rather not essential and all connectivity indices can be accepted in principle; (b) a variable index is good, but has no preference over the solvation one; and (c) these results are rather predictable due to a great structural similarity and uniformity of the data set and, even, the type of property (A-type of property in accordance with ref 9) and structural similarity of the data set.

What should be done in a situation, where one is faced with a number of rather good QSAR models? In our opinion, the key answer must be provided by the estimation of predictive power of the models.

Before proceeding further, let us first cite the following. The reviewer of the paper[12] raised the following comment: "Why not to consider more extensive study on a larger set of data to strengthen the case?" (cited in ref 12). In polemic with the referee comment, the authors replied the following: "We gave here the results for *smaller* sulfides... If one is interested in predicting the boiling point of *smaller* sulfides, why does one need information on compounds that are *twice* its size?".

The key word in this statement, related to our discussion, is "*predicting*". In other words, this statement means that the "smaller" set of 21 sulfides must produce the QSAR

Table 3. Comparison of Experimental and Predicted Boiling Points for the QSAR Models of 21 Sulfides

no.	compound	bp exptl	bp calcd using $^1\chi$	$^1\chi^s$	$^1\chi^f$
22	methyl s-butyl sulfide	114.5	111.6	112.5	111.8
	bp(exptl) − bp(calcd)		−2.9	−2.0	−2.7
23	methyl *tert*-amyl sulfide	128.3	127.5	127.9	127.3
	bp(exptl) − bp(calcd)		−0.8	−0.4	−1.0
25	methyl 1,2-dimethylpropyl sulfide	133.0	130.5	132.6	131.9
	bp(exptl) − bp(calcd)		−2.5	−0.4	−1.1

correlation "of high quality" which, in turn, can be used for *prediction*, i.e., for "a priori" estimation of boiling points of sulfides beyond the list of 21.

Hence, the next question may be immediately raised: how many "smaller" sulfides in the range from C_2H_6S to $C_6H_{14}S$ exist, which boiling points are the object of prediction.

We have performed an exhaustive generation of formulas of isomeric sulfides from C_2H_6S to $C_8H_{18}S$ using the structural generator SMOG.[35,36] The calculated numbers of isomers were the following: C_2H_6S − 1, C_3H_8S − 1, $C_4H_{10}S$ − 3, $C_5H_{12}S$ − 6, $C_6H_{14}S$ − 15, $C_7H_{16}S$ − 33, and $C_8H_{18}S$ − 82. Thus, the structural generator revealed the existence of exactly 26 sulfides in the range of molecular formulas, which used to be classified as "*smaller*" sulfides. In other words, if one talks about prediction, one must deal with *five* more structures and the target for "construction of high-quality structure-property-activity relationship"[12] is just five (and only five!) sulfides. These five sulfides **22**−**26** are also explicitly shown in Table 1.

It is quite natural to use these structures for a real test of quality of QSAR correlations using different descriptors. We were able to find the boiling point data for only three of them;[37] these data are shown in Table 1 for sulfides **22**, **23**, and **25**. We have used them as the external test set, and statistics for these data are shown in Table 2. Comparison of experimental and predicted boiling points is given explicitly in Table 3.

First, these data permit one immediately to reject the Hosoya,[38] Wiener, and S_w indices which means that these regressions are exclusively descriptive models. Second, the predicted data for solvation index $^1\chi^s$ are slightly preferable as compared with both $^1\chi$ and, what is remarkable, variable index $^1\chi^f$.

Thus, these data show that (a) selection of solvation index is justified not only by structural reasons of the data set (vide supra) but also by a real prediction; (b) the preference for application of variable index $^1\chi^f$, declared in the paper in ref 12, at least for a considered particular data set is illusive; and (c) consideration of exclusively statistical parameters without testing, as it was made in ref 12, does not permit selection of a real high quality *predictive* QSAR model.

Now return to the above-mentioned comment concerning the data set size. We decided really "to strengthen the case" and performed QSAR modeling using 38 sulfides (Table 1). The data for sulfides **27**−**43** were taken from ref 13, and this set was supplemented by two novel sulfides **44** and **45** from ref 40.

First, we performed a big extrapolation using the compounds **22**, **23**, **25**, and **27**−**45** as an extensive large test set for QSAR models, obtained above with a training set of 21 sulfides (Table 2). These data show that prediction in this case is worse but still quite reasonable. Again, the data for the solvation index $^1\chi^s$ are preferable as compared with both $^1\chi$ and variable index $^1\chi^f$.

Second, we have constructed the QSAR models for a training set of 38 sulfides with two test sets: of three (compounds **22**, **23**, **25**) and of five (compounds **22**, **23**, **25**, **45**, and **46**) sulfides; these data are shown in Table 4. Again we came to the analogous conclusion that the solvation index $^1\chi^s$ is preferable as compared with both the $^1\chi$ and the variable index $^1\chi^f$, especially for prediction data. In general, the data for a larger set are a little more diffuse. This is understandable, taking into account that the boiling points, especially at higher temperatures, are not precise due to many reasons such as partial decomposition at elevated temperatures, experimental errors, obtaining of data at low pressure and extrapolation to a normal one, etc.[41] (For an excellent discussion about the crucial influence of the quality of boiling point data see ref 11.).

CONCLUSION

We investigated several structure-boiling point models of sulfides. The results obtained lead us to the following conclusions.

First, the high-quality regressions[12] are not exactly equal to high-quality *predictive* QSAR models. The important conscience is simple: omitting a test set is methodologically invalid and does not permit one to find the best QSAR models.

Second, the logical selection of an index depending on the structural features of the training set of compounds can be helpful. For instance, we presented above the argument

Table 4. QSAR Models (bp = $ax + b$) for Boiling Points of a Larger Set of 38 Alkyl Sulfides (Compounds **1**−**21**, **27**−**43**) and Their Statistical Parameters

index	W	S_w	Z	$^1\chi$	$^1\chi^s$	$^1\chi^f$
a	0.9931	14.65	2.166	47.23	49.60	55.30
b	75.45	2.994	85.24	−21.88	−55.18	−48.02
R^2	0.8823	0.9193	0.8203	0.9774	0.9851	0.9854
s	11.7	9.7	14.4	5.1	4.2	4.1
F	270	410	164	1559	2373	2431
max. err	42.1	28.2	54.4	11.6	12.7	11.2
			Test Set: Compounds **22, 23, 25**			
RMS	9.9	3.4	8.2	3.8	2.6	3.3
max. pred. err	11.9	5.5	10.9	4.7	2.8	3.5
			Test Set: Compounds **22, 23, 25, 44, 45**			
RMS	9. 9	8.4	6.5	6.2	2.9	3.9
max. pred. err	12.5	15.8	10.9	11.0	4.7	6.4

for the selection of the solvation index, $^1\chi^s$, and it permits one to provide a fairly good prediction for the case of simple sulfides, quite comparable or even preferable with a three-parameter variable index, $^1\chi^f$.

One more general comment concerning the selection of the descriptors was raised by the reviewer: the problem of selection vs "interpretability" of the descriptors was pointed out. Probably the general tendency has to be as follows: when one has to deal with the *predictive* QSAR, one should select those descriptors, which provide the best quality of the prediction. However for the *descriptive* QSAR the selection of the descriptors may also be determined by our "understanding" of the correlation. In other words, one may prefer to select for this case the descriptors, having either physicochemical meaning or being subgraphs and, hence, having chemical structural sense.

Third, the Randić index[32] and the family of derived connectivity indices have found very wide applications in QSAR.[42,45] The principle of construction of variable indices is out of the question, and these indices have potential for QSAR modeling as multiparametric ones.[46] However, the scope and limitation of their application still needs to be revealed. Another problem arises with an interpretation of the data: the limit of variation of adjustable parameters for variable connectivity indices as well as a chemical sense of these variations is still unclear and needs to be carefully studied.

In this connection we also would like to point out that a previously developed "neural network device"[47] uses only structural formulas as the input data. That means that this device can in principle model the variability of indices based on connectivity. Unfortunately, this device requires large data sets to be properly trained. In the application to the sulfides **1**−**21** we usually obtained overtrained models; anyhow this idea is worthy of future consideration.

APPENDIX

The index based on vertex weights is computed in the following way.[27] For a hydrogen depleted molecular graph the distance matrix is calculated, and on the first step to each *i*th vertex the initial weight equal to 1 is assigned

$$w_i^{(0)} = 1,\ i = 1,..., N$$

where N is the number of vertices in a graph (number of non-hydrogen atoms).

Then for each vertex the modified weight is calculated

$$w_i'^{(0)} = w_i^{(0)} + \sum_{\substack{j=1 \\ j\neq i}}^{N} w_j^{(0)}/d_{ij}$$

where d_{ij} is the shortest distance from *i*th to *j*th vertex. Among the obtained values $w_i'^{(0)}$ the minimal value $M_w^{(0)}$ is evaluated, and then new vertex weights are calculated in the following way:

$$w_i^{(1)} = w_i'^{(0)}/M_w^{(0)}$$

After $k + 1$ iterations

$$w_i'^{(k)} = w_i^{(k)} + \sum_{\substack{j=1 \\ j\neq i}}^{N} w_j^{(k)}/d_{ij}$$

$$w_i^{(k+1)} = w_i'^{(k)}/M_w^{(k)}$$

The iterations are repeated until the value of the sum of vertex weights w_i for all vertices in a molecular graph on the consecutive iterations does not change for more than ϵ. Usually 7−10 iterations are enough to reach $\epsilon = 0.001$ for graphs having 10−25 vertices. Several descriptors can be constructed from these data; here we used the sum of vertex weights in a molecular graph as a topological index S_w.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington DC, 1995.

(2) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, Germany, 1993.

(3) Silipo, C. *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; *Pharmacochemistry Library*; Timmerman, H., Eds.; Elsevier: 1991; Vol. 16.

(4) Gordeeva, E. V.; Molchanova, M. S.; Zefirov, N. S. General Methodology and Computer Program for the Exhaustive Restoring of Chemical Structures by Molecular Connectivity Indexes. Solution of the Inverse Problem in QSAR/QSPR. *Tetrahedron Comput. Methodol.* **1990**, *3*, 389−415.

(5) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. I.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 630−634.

(6) Zefirov, N. S.; Palyulin, V. A.; Radchenko, E. V. Problem of generation of structures with predetermined properties. Solution of inverse QSAR problem for Balaban's centric index (Russ.). *Dokl. Akad. Nauk.* **1993**, *316*, 921−924 (*Chem. Abstr. 115*, 48463).

(7) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. I.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR Studies for the Case of Equations, Containing any Topological Indices (Russ.). *Dokl. Akad. Nauk.* **1996**, *346*, 497−500 (*Chem. Abstr. 125*, 166953).

(8) Skvortsova, M. I.; Slovokhotova, O. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR Studies for the Case of Informational Topological Indices (Russ.). *Dokl. Akad. Nauk.* **1997**, *357*, 72−74 (*Chem. Abstr. 128*, 216988).

(9) Hosoya, H.; Gotoh, M.; Murakami, M.; Ikeda, S. Topological Index and Thermodynamical Properties. 5. How Can We Explain the Topological Dependency of Thermodynamic Properties of Alkanes with Topology of Graphs? *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 192−196.

(10) See, however, excellent discussion for QSAR of boiling points of structurally extended set of saturated hydrocarbons in ref 11.

(11) Rücker, G.; Rücker, C. On Topological Indices, Boiling Points, and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788−802.

(12) Randić, M.; Basak, S. Construction of High-Quality Structure−Property-Activity Regression: The Boiling Points of Sulfides. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899−905.

(13) 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.

(14) The structure **21**, entry 21, Figure 1[12] is incorrect, and we have used the structure of 1-ethylpropyl methyl sulfide (entry 75, Table 2, ref 13). However the calculated values for connectivity indices are correct.

(15) Antipin, I. S.; Arslanov, N. A.; Palyulin, V. A.; Konovalov, A. I.; Zefirov, N. S. Solvation Topological Index. Topological Description of Dispersion Interactions (Russ.). *Dokl. Akad. Nauk. SSSR* **1991**, *316*, 925−927 (*Chem. Abstr. 115*, 91390).

(16) Antipin, I. S.; Arslanov, N. A.; Palyulin, V. A.; Konovalov, A. I.; Zefirov, N. S. Prognosis of Enthalpy of Nonspecific Solvation of Organic Nonelectrolytes (Russ.). *Dokl. Akad. Nauk.* **1993**, *331*, 173−176 (*Chem. Abstr. 120*, 133743).

(17) Barysz, M.; Jashari, G.; Lall, R. S.; Srivastava, V. K.; Trinajstic, N. On the Distance Matrix of Molecules Containing Heteroatoms. In *Applications of Chemical Topology and Graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; pp 222−227.

(18) Balaban, A. T. Chemical Graphs. 48. Topological Index ***J*** for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115−122.
(19) 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.
(20) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Design of Topological Indices. Part 10. Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395−401.
(21) The calculations were made using the QSAR program complex EMMA,[22−28] which calculates several hundreds of different descriptors.
(22) 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.* **1993**, *330*, 468−472 (*Chem. Abstr. 120*, 54070).
(23) Sukhachev, D. V.; Pivina, T. S.; Shlyapochnikov, V. A.; Petrov, E. A.; Palyulin, V. A.; Zefirov, N. S. Study of Quantitative Structure-impact Sensitivity Relations for Organic Polynitrocompounds. *Dokl. Akad. Nauk.* **1993**, *328*, 188−189 (*Chem. Abstr. 118*, 257697).
(24) 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).
(25) 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. 2. Liquid State. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1995**, 1657−1660 (*Chem. Abstr. 124*, 260219).
(26) 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. 3. Solid State. *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1995**, 1661−1665 (*Chem. Abstr. 124*, 260220).
(27) 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.* **1992**, *324*, 1019−1022 (*Chem. Abstr. 118*, 38223).
(28) 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.* **1992**, *327*, 504−508 (*Chem. Abstr. 118*, 182798).
(29) Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.
(30) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(31) Tratch, S. S.; Stankevitch, M. I.; Zefirov, N. S. Combinatorial Models and Algorithms in Chemistry. The Expanded Wiener Number − a Novel Topological Index. *J. Comput. Chem.* **1990**, *11*, 899−908.
(32) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(33) Standard deviation for $^1\chi$ was found equal to 3.31 but not to 2.701 as in Table 1.[12] M. Randić (personal communication) clarified that Table 1 is based, in fact, on $n = 19$ but not on 21 compounds, because two compounds, **1** and **20**, were eliminated as outliers.
(34) One has to be careful in recalculating the generalized flexible connectivity indices using Table 4, ref 12, because the formulas of the entries 4, 9, 14, 15, 19, and 21 contain misprints.
(35) Molchanova, M. S.; Shcherbukhin, V. V.; Zefirov, N. S. Computer Generation of Molecular Structures by SMOG Program. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 888−899.
(36) Molchanova, M. S.; Zefirov, N. S. Irredundant Generation of Isomeric Molecular Structures with some Known Fragments. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 8−22.
(37) 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.
(38) The correlations using Hosoya index, Z, are rather rare and observed usually for properties of small homogeneous series of hydrocarbons[9] (cf. ref 11). We obtained very good result using lnZ ($R^2 = 0.9900$; $s = 2.87$; $F = 1886$). The prediction data for lnZ are also very good (**22**, 113.5; **23**, 128.6; **25**, 131.8). Moreover, a switch to ln scale for $^1\chi^s$ gives an even better prediction (**22**, 114.8; **23**, 128.5; **25**, 132.4). This situation may reflect an intrinsic slight nonlinearity of correlating data for boiling points, especially for lower homologues. We have observed this phenomenon in some other correlations of boiling point sets. In fact, the paper[12] also demonstrates a better correlation for nonlinear (quadratic) regression. Better correlation for logarithmic scale was even discussed in ref 39, and this point probably needs more careful investigation.
(39) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, *λ*. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(40) Golovnya, R. V.; Misharina, T. A.; Garbuzov, V. G. Gas-chromatografic Characteristic of Sulfur Containing Compounds. 6. Di-iso-Alifatic sulfides (Russ.). *Izv. Akad. Nauk SSSR (Bull. Acad. Sci. USSR)* **1979**, 1029−1032.
(41) The discrepancy of boiling points data for sulfides in the range of 1−4° was discussed in ref 13 (one example showed as much as 10°!). Compare also, for instance, the experimental boiling points for isopropyl isobutyl sulfide, **30**, equal 145.0°[13] and 142.0°.[40]
(42) Clarification of a meaning of connectivity indices see ref 43; index range see ref 44.
(43) Kier, L. B.; Hall, L. H. Intermolecular Accessibility: The Meaning of Molecular Connectivity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792−795.
(44) Gutman, I.; Araujo, O.; Morales, D. A. Bounds for the Randić Connectivity Index. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 593−598.
(45) Stankevich, I. V.; Stankevich, M. I.; Zefirov, N. S. Topological indices in organic chemistry (Russ.). Usp. Khim. *(Russ. Chem. Rev.)* **1988**, *57*, 337−366 (*Chem. Abstr. 125*, 166953).
(46) We mean a general case: the number of extra variables depends on a number of atom types.
(47) 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.

CI0001637