

# Construction of High-Quality Structure–Property–Activity Regressions: The Boiling Points of Sulfides

Milan Randić<sup>\*,†,‡</sup> and Subhash C. Basak<sup>§,⊥</sup>

Department of Mathematics and Computer Science, Drake University, Des Moines, Iowa 50311 and  
Natural Resources Research Institute, The University of Minnesota, 5013 Miller Trunk Highway,  
Duluth, Minnesota 55811

Received September 9, 1999

Instead of using the standard molecular descriptors (topological indices) for regression analysis, which are numerically fully determined once a molecule is selected, we outline the use of variable molecular descriptors that are modified during the search for the best regression. The approach is illustrated using boiling points of sulfides. We have transformed the connectivity index  ${}^1\chi$  into a function of two variables ( $x$ ,  $y$ ) which differentiate carbon and sulfur atoms. The optimal values of the variables ( $x$ ,  $y$ ) were determined by minimizing the standard error of the regression. With the values  $x = +0.25$  and  $y = -0.95$  for carbon and sulfur, respectively, we have obtained a regression based on a single descriptor and a standard error of 1.8 °C. With elimination of two outliers (having a deviation of about 4 °C) the standard error is reduced to a remarkable 1.3 °C.

## INTRODUCTION

The past decade has witnessed two important developments of multivariate regression analysis, MRA, relevant for quantitative structure–property–activity relationship, QSAR: (1) expansion of mathematical structural descriptors for characterization of molecular structure;<sup>1–5</sup> (2) construction of orthogonal molecular descriptors<sup>6–12</sup> which result in stable regression equations. The first, which is of interest when better regressions are sought, is rather conspicuous, while the second, which is important for interpretation of the results of such studies, remains not yet sufficiently widely appreciated.

In this paper we will address the problem of construction of high-quality regressions (HQR). With hundreds of descriptors available<sup>13–15</sup> the questions to consider are as follows: (1) How should an optimal set of descriptors be chosen from a large number of available descriptors? (2) How should one choose between regressions of seemingly similar quality? (3) How unique are regression results? (4) Are there important structural elements missed by the descriptors used? (5) How complete is the space spanned by molecular descriptors for the structure–property–activity studies? (6) Do we need additional molecular descriptors?

## HIGH-QUALITY REGRESSIONS

The standard error in most correlations still does not approach the experimental error of measurements. How realistic is it to hope to arrive at this goal? As we will show, HQR, in which the standard error has been dramatically reduced in comparison with traditional approaches using the same number of descriptors, can be derived with a new kind

**Table 1.** Standard error for the Boiling Points of Smaller Sulfides ( $n = 21$  Compounds) for Selection of Descriptors

descriptors	standard error	descriptors	standard error
$\chi$ , $J$	2.001	$\chi$	<b>2.701</b>
$\chi$ , $n$	2.550	$n$ , $J$	2.748
$\chi$ , $P$	2.560	$n$ , $p_2/w_2$	2.981
$\chi$ , $W$	2.667	$J$ , $W$	4.808
$\chi$ , $p_2/w_2$	2.692	$W$ , $P$	5.109

of molecular descriptors which involve variability that allows one to optimize the descriptors and minimize the standard error of regression.

In Table 1 we illustrate the standard errors for correlations of the boiling points of smaller sulfides (shown in Figure 1) using a selection of molecular descriptors. When the connectivity index<sup>16</sup> is used alone, we find the standard error of the regression is 2.70 °C, as shown in the middle of Table 1. When the connectivity index is combined with Balaban's  $J$  index,<sup>17</sup> the standard error is further reduced to 2.00 °C. Other descriptors, viz.,  $n$ , the number of non-hydrogen atoms,  $P_3$ , the number of paths of length 3,  $W$ , the Wiener index,<sup>18</sup> and the  $p_k/w_k$ , path/walk quotients,<sup>19</sup> give only a minor improvement for the standard error over that based on  ${}^1\chi$  used alone. In contrast other combinations of molecular descriptors (listed in the right part of Table 1) do not give satisfactory results. The standard error in such combinations is worse than the standard error when the connectivity index is used as a single descriptor, which well-illustrates the importance of the proper selection of molecular descriptors.

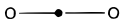

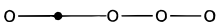
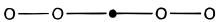
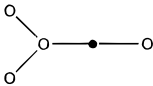
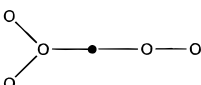

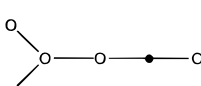

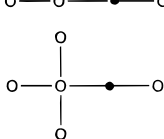
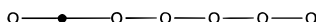
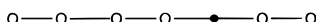

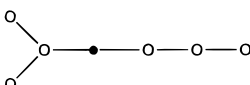
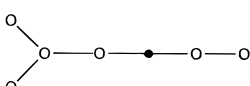
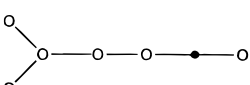
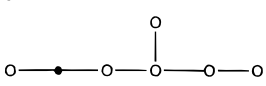
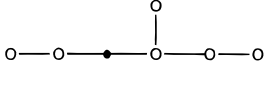

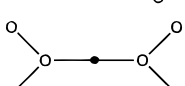
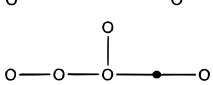
The compounds considered here were among 45 saturated acyclic compounds possessing divalent sulfur atoms for which Balaban et al.<sup>20</sup> found reliable literature data. We took all compounds having six or fewer carbon atoms, a total of 21, and have recalculated the regressions for only these smaller sulfides. The study of Balaban and co-workers considered a broader class of compounds: 185 saturated

<sup>†</sup> Drake University and The University of Minnesota.

<sup>‡</sup> FAX (home): 515 292 8629.

<sup>§</sup> The University of Minnesota.

<sup>⊥</sup> E-mail: sbasak@nrri.umn.edu.

1		37.3
2		66.6
3		95.5
4		92.0
5		84.4
6		107.4
7		123.2
8		112.5
9		118.5
10		101.5
11		145.0
12		144.2
13		142.8
14		132.0
15		134.2
16		137.0
17		139.0
18		133.6
19		120.4
20		120.0
21		137.0

**Figure 1.** Molecular graphs of smaller sulfides and their boiling points. The sulfur atoms are shown as a filled circle.

acyclic compounds possessing divalent oxygen or sulfur atoms, and devoid of hydrogen bonding, having 11 or less non-hydrogen atoms. Their purpose was as follows: (i) to explore the role of heteroatoms within acyclic skeletons in determining a measured molecular property (boiling points); (ii) to show that topological descriptors can satisfactorily account for the observed relative magnitudes of the property; and (iii) to derive structure–property regressions that may be useful for predicting boiling points of unknown compounds.

Our objectives are the same, but our philosophy in this particular study is somewhat different: Rather than considering a large set of mixed compounds (alkanes, ethers, diethers, acetals, and peroxides as well as their sulfur analogues: sulfides (thioethers), bis-sulfides, thioacetals, and disulfides), which allows one to use several molecular descriptors and still maintain high statistical significance for the correlation, we decided to use only structurally closely related compounds. In particular, we excluded bis-sulfides and disulfides because of the presence of S–S linkage that is absent in sulfides. This has reduced the pool of the compounds considerably, which limits the number of descriptors that one should use in analyzing the data. By homogenizing the sample of the compounds to be examined, as we will see, we can achieve a very high quality regression result using a *single* descriptor.

As we see from Table 1, apparently it is difficult to reduce the standard error for the boiling points of sulfides below 2.5 °C. Among the combinations listed in Table 1, only Balaban's *J* reduced the standard error below 2.5 °C. This may not be surprising because all descriptors of Table 1 except *J* do not differentiate sulfur and carbon atoms. Hence, 2.5 °C may well be the limit that such models can attain. The experimental boiling points for butylmethyl sulfide (7) and ethylpropyl sulfide (9), 123.2 and 118.5 °C, respectively, differ by almost 5 °C. If we overlook the difference between sulfur and carbon, both these structures have the same molecular graph. The same is true for ethylisopropyl sulfide (6) and isobutylmethyl sulfide (8), with the boiling points 107.4 and 112.5 °C, respectively. Hence, the simple connectivity index and other topological indices that do not discriminate heteroatoms can at best approach the standard error of about 2.5 °C.

Observe that the descriptors listed in Table 1 are of quite distinct structural origin and thus do not duplicate one another. However, many of such indices, even when combined (the right part of Table 1), apparently lack flexibility to represent the data with desirable accuracy. Using descriptors that differentiate heteroatoms, we reach a standard error of about 2 °C. The question to consider is as follows: Can the standard error of 2 °C obtained using  $^1\chi$  and *J* be further dramatically reduced? Have we reached the limit for correlating the boiling points of sulfides? Is it that the residual of the molecular property considered cannot be described by any of the available structural descriptors?

#### FLEXIBLE MOLECULAR DESCRIPTORS

In order to develop a high-quality regression, we not only need new descriptors but we need a *new kind* of molecular descriptors that have the flexibility to adjust to the variability that different molecules may show. One such descriptor has been introduced in the multiple regression analysis 10 years ago,<sup>21,22</sup> but apparently has been mostly overlooked. That novelty can be ignored or overlooked has already been well-illustrated by the Wiener index *W*, which waited two decades to be resurrected. In order to not repeat that history, we undertook a concerted effort to illustrate properties of variable descriptors, and the variable connectivity index, in particular.<sup>23–26</sup> The variable connectivity index represents an important and distinct generalization of the connectivity index  $^1\chi$  since it offers a flexibility that traditional topological indices, all several hundred of them, have been lacking.

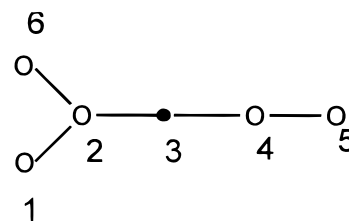
We propose here a special symbol,  ${}^1\chi^f$ , for the flexible connectivity, index which is to be outlined shortly. The original connectivity index  ${}^1\chi$  (named so by Kier et al.<sup>27</sup>), proposed by Randic,<sup>16</sup> used a fixed number as entries in the weighting algorithm  $1/(pq)^{1/2}$  for the contribution of a bond having  $p$  and  $q$  neighbors. The higher order connectivity indices,  ${}^m\chi$ ,<sup>28</sup> were defined analogously using paths of length  $m$ , for  $m = 2, 3, \dots$ . The bonding connectivity indices,  ${}^1\chi^b$ , were considered by Basak and Magnuson<sup>29</sup> on the basis of weights equal to the number of bonds of an atom: 1 for a single bond, 2 for a double bond, and 3 for a triple bond. The valence connectivity indices,  ${}^1\chi^v$ , developed by Kier and Hall,<sup>30</sup> use the difference in valence electrons and the number of hydrogen atoms to modify the valence parameter for heteroatoms. Finally "edge connectivity" indices were recently tested using bond adjacency rather than vertex adjacency in construction of the modified connectivity indices.<sup>31</sup>

All the above indices, except  ${}^1\chi^f$ , are based on fixed weights determined by the connectivity of the molecular graph model used. In our view, a better strategy is to introduce weights that make descriptors "flexible", so not only that atoms of different type can adjust their weights in order to yield an optimal characterization of a molecule for a particular property but that they may change values when different properties of the same set of molecules are considered. In general, for a molecule with  $n$  different types of atoms,  $x_1, x_2, \dots, x_n$ , one can have  $n$  different weights  $x_i$  ( $i = 1, 2, \dots, n$ ); hence, the flexible connectivity index  ${}^1\chi^f$  becomes a function of  $n$  variables. In the case of sulfides, we consider two variables, the weights of carbon and sulfur atoms. In the case of natural amino acids there are four kinds of atoms: carbon, oxygen, nitrogen, and sulfur; hence, in this case flexible connectivity indices  ${}^1\chi^f$  imply optimization of four variables.<sup>24</sup> Even if there are no heteroatoms, variable weights can improve regressions visibly.<sup>25</sup>

It should be noted that while the special types of connectivity indices, viz.,  ${}^m\chi$ ,  ${}^m\chi^b$ , and  ${}^m\chi^v$  indices, explore only local regions of the parameter space, the  ${}^m\chi^f$  indices are capable of exploring the full potential of the parameter space generated by the presence of heteroatoms in a molecule. The previously mentioned simple connectivity indices and valence connectivity indices can be viewed as a special case of the more general flexible indices  ${}^m\chi^f$ . Consequently, the flexible indices  ${}^m\chi^f$  are expected to be more powerful in predicting molecular properties and biological activities.

Besides the weighted connectivity indices,<sup>21–26</sup> many other topological indices, e.g. the weighted paths  $p_k^f$ ,<sup>32–34</sup> the weighted walks,  $w_k^f$ , the weighted Hosoya index  $Z^f$ , the weighted Wiener index  $W^f$ , and the weighted Balaban index  $J^f$ , can be generalized in a similar way.<sup>35</sup> Except for a half-dozen papers of the present authors,<sup>21–26,32–34</sup> use of variable molecular descriptors is in its infancy.

Dramatic improvement in the quality of regressions was obtained by using variable connectivity indices. For example, by introducing a variable parameter  $x$  for chlorine in clonidine and clonidine-like imidazolidines (2-(arylimino)-imidazolidines),<sup>21</sup> the value  $x = -0.20$  for chlorine produces a regression which, with three weighted connectivity indices, gave better results for the set of clonidine compounds as compared to five descriptors used in a traditional QSAR.<sup>36</sup>



**Figure 2.** Molecular graph of ethyl isopropyl sulfide and the corresponding numbering of atoms used in Table 2.

**Table 2.** Adjacency Matrix and Modified Adjacency Matrix for Ethyl Isopropyl Sulfide

adjacency matrix							row sum
	1	2	3	4	5	6	
1	0	1	0	0	0	0	1
2	1	0	1	0	0	0	2
3	0	1	0	1	0	0	2
4	0	0	1	0	1	1	3
5	0	0	0	1	0	0	1
6	0	0	0	1	0	0	1

modified adjacency matrix							row sum
	1	2	3	4	5	6	
1	$x$	1	0	0	0	0	$1+x$
2	1	$x$	1	0	0	0	$2+x$
3	0	1	$y$	1	0	0	$2+y$
4	0	0	1	$x$	1	1	$3+x$
5	0	0	0	1	$x$	0	$1+x$
6	0	0	0	1	0	$x$	$1+x$

This result is particularly striking for this data set, because there are two extreme potency values which would be expected to give much trouble in cross-validation. Use of two variables that differentiate carbon and oxygen in alcohols, with  $x = +1.5$  and  $y = -0.85$ , respectively, reduced the standard error of 7 °C, obtained using the simple connectivity index that does not differentiate carbon and oxygen atoms, to 3.5 °C.<sup>22</sup> In the case of amines, the standard error of 3.48 °C for the boiling point model when  ${}^1\chi$  is used has been reduced to 1.91 °C with  $x = +1.25$  and  $y = -0.65$ .<sup>23</sup> The standard error for a quadratic regression using the connectivity index for the boiling points of smaller alkanes is 2.98 °C. When  $x = +0.65$  is introduced as a weight, not only is  $s = 2.48$  obtained, a reduction by a half-degree Celsius, but higher precision allowed the recognition of an outlier (with an error of over 6 °C), which, when eliminated, further reduced the standard error to an impressive 1.57 °C.<sup>25</sup>

#### OPTIMAL DESCRIPTORS FOR SULFUR

We will examine the correlation of the boiling points for sulfides of Figure 1 using functional molecular descriptors and will illustrate the use of a variable connectivity index by considering ethyl isopropyl sulfide (shown in Figure 2 with the numbering of the atoms used). The adjacency matrix and the modified adjacency matrix of ethyl isopropyl sulfide are illustrated in Table 2. If we assume  $x = 0$  and  $y = 0$ , we obtain the usual adjacency matrix of a graph from the row sums of which the simple connectivity index can be directly computed. To obtain the bond contribution for  ${}^1\chi$ , we use the algorithm  $1/(pq)^{1/2}$ . Here  $m$  and  $n$  are the respective valences as obtained from the row sums for atoms  $m$  and  $n$  forming the bond  $(p, q)$ . When  $x \neq 0$  and  $y \neq 0$ , the corresponding row sums are modified, and instead of the

**Table 3.** Modified Connectivity Index  ${}^1\chi$  for Ethyl Isopropyl Sulfide with Different Choices of  $x$  and  $y$ 

$x$	$y$	${}^1\chi(x, y)$	$x$	$y$	${}^1\chi(x, y)$
0	-1.00	4.392 51	+0.25	-0.95	2.780 49
0	-1.20	3.297 87	0	0	2.770 06
0	-1.00	3.146 26	+0.25	-0.90	2.753 09
0	-0.95	3.115 31	0	+0.50	2.674 17
0	-0.90	3.086 49	+0.50	-1.00	2.556 25
0	-0.75	3.010 66	+0.50	-0.95	2.528 12
0	-0.50	2.910 56	+1.00	-1.00	2.192 71
0	-0.25	2.832 77	+2.00	-1.00	1.752 29
+0.25	-1.00	2.809 93			

fixed valences  $p$ ,  $q$ , we have the variable valence  $(p + x)$ ,  $(q + x)$ , or  $(q + y)$ , depending on the kind of atoms involved. Thus instead of the simple ("fixed") connectivity index  ${}^1\chi = 1/\sqrt{2} + 1/2 + 1/\sqrt{6} + 2/\sqrt{3}$ , we have the variable connectivity index given as a function of two variables:

$${}^1\chi(x, y) = 1/\{(1+x)(2+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(3+x)(2+y)\}^{1/2} + 2/\{(1+x)(3+x)\}^{1/2}$$

In Table 3 we listed selected values of the variable  ${}^1\chi$  molecular descriptor for ethyl isopropyl sulfide. As we see, the flexible descriptor is sensitive on the choice of the values for  $x$  and  $y$ . For a fixed value of  $x$  (carbon atom), as  $y$  decreases and approaches  $-1$ , the magnitudes of the modified connectivity index increase. Similarly for a fixed value of  $y$  as  $x$  increases the magnitude of the modified connectivity index decreases. An increase and a decrease of the modified index is not so important as is the change of the relative magnitudes of the indices for different molecules.

In Table 4 we have listed the expressions for the modified connectivity indices for the set of  $n = 21$  sulfides. In order to illustrate the flexibility of these generalized connectivity indices in Table 5, we listed for the selected values of  $x$  and  $y$  the numerical values for the variable connectivity indices. Even though for most of the structures the numerical magnitudes have not reversed the relative magnitudes, they altered the magnitudes of the indices for different molecules sufficiently to influence the quality of the regression dramatically. The ratios of the magnitudes of descriptors for different molecules are important for MRA, and these do change. Consider isopropyl propyl sulfide (**14**) and ethyl isobutyl sulfide (**15**) with the boiling points 132.0 and 134.2 °C, respectively. As we can see from Table 5 when  $x = -1/2$ , and  $y = -1$ , the modified connectivity indices are as follows: 5.059 17 and 5.092 95, giving the quotient 0.9934. However, when  $x = +1/2$  and  $y = -1$  the modified connectivities are as follows: 2.956 25 and 2.992 24, and the quotient decreases to 0.9880. These changes may appear small; however, they are sufficient enough to influence the standard error and make one alternative better than the other. When such changes are summed for all molecules, considerable improvement in the overall standard error is possible.

In Table 6 we show the standard error as a function of the parameters  $x$ ,  $y$ , assuming a quadratic regression using  $n = 19$  compounds. We excluded two structures, ethyl butyl sulfide **12** and diisopropyl sulfide **20**, to be discussed later. Using the simple connectivity index, the (0, 0) point in Table 6, the standard error is quite respectable 2.71 °C. Nevertheless this is about twice the magnitude of typical experimental

errors reported for boiling points of organic compounds (1–1.5 °C). By keeping  $x$  constant and varying  $y$ , we see a dramatic reduction of the standard error as we approach the  $y = -1$  limit. The standard error for  $x = 0$  and  $y = -1$  is about 1.5 °C smaller than the initial value ( $x = y = 0$ ). With a further change of both parameters  $x$  and  $y$ , we find the minimum standard error of 1.326 °C (when  $x = +0.25$  and  $y = -0.95$ ). This is less than half of the initial standard error characterizing the "inflexible" connectivity index.

## OUTLIERS

Mathematical descriptors, if correctly calculated, are error-free. Hence, if in a correlation between an experimental quantity and mathematical descriptors of one or more points show larger deviation from the regression curve, this can mean two things: Either (1) some experimental data used are in *error* or (2) the descriptors used *fail* to capture some relevant structural feature present in some (and absent in other) molecules.

Whatever is the reason for the departure of a point from the regression line, one can consider such a point as an outlier if the departure from the correlation is more than twice the standard error. In Figure 3 we show the quadratic correlation for sulfides, and in Table 7 we listed the computed boiling point and the residue. As we see from Table 7 ethyl butyl sulfide and diisopropyl sulfide show large departures from the regression. In Table 8 are given the regression equations and the associated statistical parameters for all  $n = 21$  sulfides as well as for the cases  $n = 19$  sulfides where two outliers have been removed respectively from the set considered.

By eliminating the apparent outliers (**12** and **20**), one substantially reduces the standard error for the quadratic model, as can be seen from the bottom part of Table 8. The standard error for the regression when  $n = 19$  reaches the respectable value of 1.33 °C and the correlation coefficient and the Fisher ratio have increased. This signals that the model has improved and that we were justified in eliminating the two outliers.

In Table 9 we listed the optimal connectivity indices for the sulfides considered, the experimental boiling points (BP), the calculated boiling points (BPcalc), the residual of the regression (Res), the cross-validated boiling points ( $\times$ BPcalc), and the standard error associated with cross-validation (when leaving one entry out). For the two outliers, ethyl butyl sulfide and diisopropyl sulfide, which were excluded when the regression equation was derived, we calculate for the boiling points to be 140.44 and 124.47 °C, respectively. The first of these values is about 4 °C below the reported experimental BP; the second value is almost 4.5 °C higher than the reported experimental BP. The quadratic regression without the data on the two outliers is illustrated in Figure 4.

A closer look at the last column of Table 9, which lists the standard errors associated with the cross-validated regressions, shows (with a single exception **13**, dipropyl sulfide) that the cross-validated standard errors differ about  $\pm 0.05$  °C from the standard error of the regression (when all  $n = 19$  compounds are considered). Hence, disregarding the exception which produced significantly *smaller* standard error, the constancy of the cross-validated standard errors show the robustness of this particular regression.



**Table 4.** Generalized Flexible Connectivity Indices for  $n = 21$  Sulfides (of Figure 1)

1	$2/\{(1+x)(2+y)\}^{1/2}$
2	$1/\{(1+x)(2+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
3	$1/\{(1+x)(2+x)\}^{1/2} + 1/(x+2) + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
4	$1/(x+2) + 1/\{(2+x)(2+y)\}^{1/2}$
5	$2/\{(1+x)(3+x)\}^{1/2} + 1/\{(3+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
6	$2/\{(1+x)(3+x)\}^{1/2} + 1/\{(3+x)(2+y)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+x)\}^{1/2}$
7	$1/\{(1+x)(2+x)\}^{1/2} + 2/(2+x) + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
8	$2/\{(1+x)(3+x)\}^{1/2} + 1/\{(3+x)(2+y)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
9	$2/\{(1+x)(2+x)\}^{1/2} + 2/(2+x) + 2/\{(2+x)(2+y)\}^{1/2}$
10	$3/\{(1+x)(4+x)\}^{1/2} + 1/\{(4+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
11	$1/\{(1+x)(2+x)\}^{1/2} + 3/(2+x) + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
12	$2/\{(1+x)(2+x)\}^{1/2} + 2/(2+x) + 2/\{(2+x)(2+y)\}^{1/2}$
13	$2/\{(1+x)(2+x)\}^{1/2} + 2/(2+x) + 2/\{(2+x)(2+y)\}^{1/2}$
14	$2/\{(1+x)(3+x)\}^{1/2} + 1/(2+x) + 1/\{(3+x)(2+y)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
15	$2/\{(1+x)(3+x)\}^{1/2} + 1/\{(3+x)(2+x)\}^{1/2} + 2/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
16	$2/\{(1+x)(3+x)\}^{1/2} + 1/(2+x) + 1/\{(3+x)(2+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
17	$1/\{(1+x)(2+x)\}^{1/2} + 1/\{(1+x)(3+x)\}^{1/2} + 2/\{(2+x)(3+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$
18	$2/\{(1+x)(2+x)\}^{1/2} + 1/\{(1+x)(3+x)\}^{1/2} + 1/\{(2+x)(3+x)\}^{1/2} + 1/\{(3+x)(2+y)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2}$
19	$1/\{(1+x)(2+x)\}^{1/2} + 3/\{(1+x)(3+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(4+x)(2+y)\}^{1/2}$
20	$4/\{(1+x)(3+x)\}^{1/2} + 2/\{(3+x)(2+y)\}^{1/2}$
21	$1/\{(1+x)(2+x)\}^{1/2} + 1/\{(1+x)(3+x)\}^{1/2} + 1/\{(2+x)(3+x)\}^{1/2} + 1/\{(2+x)(2+y)\}^{1/2} + 1/\{(1+x)(2+y)\}^{1/2}$

**Table 5.** Modified Connectivity Index  ${}^1\chi$  for Sulfide for a Selection of Choices of  $x$  and  $y$ 

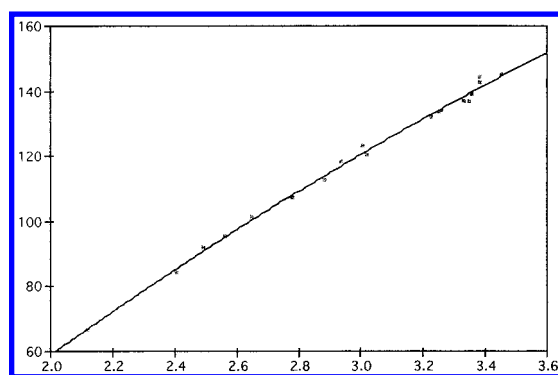
	(0, 0)	(0, -0.5)	(0, -1)	(-0.5, -1)	(+0.5, -1)	(+1, -1)
1	1.414 21	1.632 99	2.000 00	2.828 43	1.632 99	1.414 21
2	1.914 21	2.100 95	2.414 21	3.385 41	1.965 35	1.692 71
3	2.414 21	2.600 95	2.914 21	4.052 08	2.365 35	2.026 04
4	2.414 21	2.568 91	2.828 23	3.942 39	2.297 71	1.971 20
5	2.270 06	2.442 60	2.732 05	3.835 52	2.223 89	1.914 21
6	2.770 06	2.910 56	3.146 26	4.392 51	2.556 25	2.192 71
7	2.914 21	3.100 95	3.414 21	4.718 74	2.765 35	2.359 37
8	2.770 06	2.956 80	3.270 06	4.535 96	2.659 89	2.289 24
9	2.914 21	3.068 91	3.328 43	4.609 06	2.697 71	2.304 53
10	2.560 66	2.724 74	3.000 00	4.216 52	2.442 60	2.103 00
11	3.414 21	3.600 96	3.914 21	5.385 41	3.165 35	2.692 71
12	3.414 21	3.568 91	3.828 43	5.275 37	3.097 71	2.637 86
13	3.414 21	3.568 91	3.828 43	5.275 37	3.097 71	2.637 86
14	3.270 06	3.410 56	3.646 26	5.059 17	2.956 25	2.526 04
15	3.270 06	3.424 76	3.684 27	5.092 95	2.992 24	2.558 73
16	3.270 06	3.456 80	3.770 06	5.202 63	3.059 89	2.613 57
17	3.308 06	3.494 80	3.808 06	5.312 63	3.077 91	2.623 61
18	3.308 06	3.448 57	3.684 27	5.169 18	2.974 27	2.536 08
19	3.060 67	3.192 71	3.414 21	4.773 51	2.774 96	2.381 50
20	3.125 90	3.252 21	3.464 10	4.842 62	2.814 79	2.414 21
21	3.346 07	3.518 61	3.808 06	5.388 87	3.059 94	2.600 95

**Table 6.** Standard Error of the Regression for Different Choices of the Variable Parameters  $x$  and  $y$ 

	-0.5	0	+0.25	+0.50	+1	+2
+0.50		3.273				
0		2.711				
-0.25		2.363				
-0.50		1.966				
-0.75		1.558				
-0.90		<b>1.382</b>	1.347			
-0.95		1.356	1.326	1.380		
-1	2.256	1.357	1.327	1.327	1.570	2.042
-1.2		1.720				

We believe that it may be possible to further improve the regression. A close inspection of residuals shows, with very few exceptions, that all linear structures have positive residual, while all branched structures show a negative residual. This suggests the possibility for further reduction of the standard error (particularly if the exceptions are viewed as outliers). However, such refinements should be attempted when a larger set of compounds is considered in order to see if the observed trend is genuine or not.

Finally, as a warning, we should add that when using flexible descriptors, elimination of outliers may influence

**Figure 3.** 3. Quadratic regression for the boiling points of  $n = 21$  sulfides against the optimal connectivity index ( $x = +0.25$ ,  $y = -0.95$ ).**Table 7.** Calculated Boiling Points (BPcalc) and the Residual of the Regression (Res), When All  $n = 21$  Sulfides Are Considered

	BP	BPcalc	Res
1	37.3	38.44	-1.14
2	66.6	65.53	+1.07
3	95.5	94.86	+0.64
4	92.0	90.42	+1.58
5	84.4	84.81	-0.41
6	107.4	108.01	-0.61
7	123.2	121.09	+2.11
8	112.5	114.14	-1.64
9	118.5	117.14	+1.36
10	101.5	100.04	+1.46
11	145.0	144.21	+0.79
12	144.2	140.75	<b>+3.45</b>
13	142.8	140.75	+2.05
14	132.0	132.73	-0.73
15	134.2	134.52	-0.32
16	137.0	138.12	-1.12
17	139.0	139.40	-0.40
18	133.6	134.05	-0.45
19	120.4	121.82	-1.42
20	120.0	124.31	<b>-4.31</b>
21	137.0	138.94	-1.94

the optimal values for the parameters  $x$ ,  $y$ , though not necessarily dramatically.

## CONCLUDING REMARKS

Several criticisms could be raised concerning the outlined work:<sup>37</sup> Is it appropriate to refer to MRA using flexible

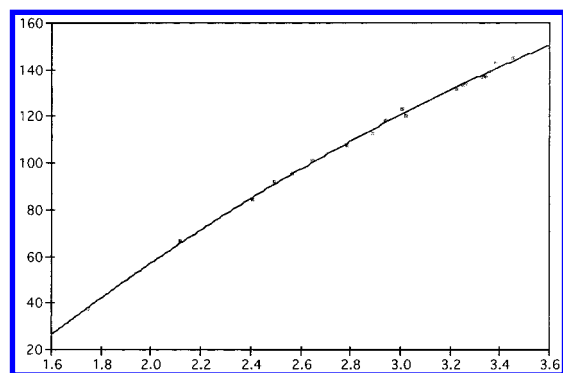
**Table 8.** Linear and Quadratic Regressions for Sulfides<sup>a</sup>

<i>n</i>	model	coeff <i>x</i>	coeff <i>x</i> <sup>2</sup>	constant	<i>r</i>	<i>s</i>	<i>F</i>
21	linear	60.1981		-61.3339	0.9959	2.61	2291
21	quadratic	102.8180	-7.8615	-117.0919	0.9981	1.83	2328
21	orthogonal	60.1981	-7.8615	-61.3339	0.9981	1.83	2328
19	linear	60.1057		-60.9916	0.9961	2.59	2180
19	quadratic	108.9647	-9.0423	-124.6847	0.9990	1.33	4192
19	orthogonal	60.1057	-9.0423	-60.9916	0.9990	1.33	4192

<sup>a</sup> The top part gives the regression equations and the statistical parameters for all *n* = 21 sulfides; the bottom part gives the equations when two outliers are excluded.

**Table 9.** Optimal Connectivity Indices for the Sulfides Considered, the Experimental Boiling Points (BP), the Calculated Boiling Points (BPcalc), the Residual of the Regression (Res), the Cross-Validated Boiling Points (×BPcalc), and the Standard Error of Cross-Validated Boiling Points

	(+0.25, -0.095)	BP	BPcalc	Res	×BPcalc	xstd error
1	1.745 75	37.3	37.98	-0.68	40.65	1.31
2	2.119 76	66.6	65.66	+0.94	65.41	1.34
3	2.564 20	95.5	95.27	+0.23	95.23	1.37
4	2.493 37	92.0	90.82	+1.18	90.60	1.33
5	2.406 48	84.4	85.17	-0.77	85.30	1.35
6	2.780 49	107.4	108.38	-0.98	108.51	1.34
7	3.008 65	123.2	121.30	+1.90	120.86	1.27
8	2.885 55	112.5	114.45	-1.95	114.66	1.26
9	2.938 21	118.5	117.41	+1.07	117.31	1.34
10	2.647 84	101.5	100.44	+1.06	100.28	1.38
11	3.453 09	145.0	143.76	+1.24	143.42	1.32
12	3.382 66	144.2				
13	3.382 66	142.8	140.44	+2.36	138.86	1.20
14	3.224 94	132.0	132.68	-0.68	132.74	1.38
15	3.259 56	134.2	134.42	-0.22	134.44	1.37
16	3.329 99	137.0	137.90	-0.90	138.01	1.35
17	3.355 50	139.0	139.14	-0.14	139.16	1.37
18	3.250 44	133.6	133.96	-0.36	134.00	1.37
19	3.021 85	120.4	122.02	-1.62	122.16	1.30
20	3.067 22	120.0				
21	3.346 37	137.0	138.69	-1.69	138.94	1.29

**Figure 4.** Quadratic regression for the boiling points of *n* = 19 sulfides against the optimal connectivity index (*x* = +0.25, *y* = -0.95). Outliers excluded **12** and **20**.

descriptors as “high-quality regression”, or should it be called “high specialty SAR”? Is one justified to arrive at low standard error by “trimming the data set and by tweaking the descriptor”? Would the model be any good to predict boiling points even for other sulfides? Is the approach general enough and sufficiently justified if we were to use QSAR models for real world problems? Why not consider more extensive study on a larger set of data to strengthen the case? What is the use of a model developed by considering a quite small, homogeneous set of compounds? Is developing a fit with standard error less than that of the experimental error (if that can be achieved) overfitting?

We respond to these question one by one. Variable connectivity indices (and related variable indices) constitute a general class of descriptors as compared to the special class of descriptors used in QSAR (e.g. indicator variables used in some QSAR, or hydrogen bonding descriptors used in CODESSA) for which the attribute “high specialty” holds. Concerning the problem of identifying outliers, these are well-defined as points that are beyond 2 standard deviations. There are no good reasons for their inclusion in the data set, despite that their departure from the regression need not be due to experimental error. Most often they are not. The occurrence of outliers may be a signal that the set of descriptors used to characterize molecules failed to characterize some special structural features which are important for outliers but not for most of other molecules in the set. A close look at outliers may help one to recognize such features, if they are not obvious. For example, correlation of the boiling points of smaller alkanes<sup>25</sup> shows only 2,2,3,3-tetramethylbutane was identified as an outlier (with deviation of over 6 °C), while the standard error was 2.48 °C. By removing this compound, standard error dropped to 2 °C. Hence, a single compound in a set of 20 was able to increase the standard error almost by 1/2 °C. Why should this compound that has *additional* structural features (significant overcrowding of methyl groups and a quaternary CC bond) absent in the rest be included if one is interested in predicting the boiling point of a compound which has no overcrowded methyl groups and no quaternary CC bond?

Smaller sulfides considered (and the same has been the case with smaller alkanes or amino acids) are molecules of similar size. To consider large selection of compounds necessarily brings the dominant role of molecular size into focus as important feature. Before we do this, we should investigate to what extent the variable weights may depend on the size of the molecule. At the moment this is an unresolved problem, which is the main reason for restricting attention to smaller sets of compounds with similar size. We should add that it is not uncommon in QSAR to consider smaller sets of compounds, often because of limited data. For example in a recent review of comparative QSAR Hansch and co-workers<sup>38,39</sup> gave results for 189 regressions in which only 33 had more than 20 compounds in the set, and 156 had less than 20 compounds, that is, less than the number of sulfides considered in this paper. If compounds are well-selected, the resulting regressions may be of interest. We gave here the results for *smaller* sulfides. If one is interested in larger sulfides, one should select those, and if one is interested in all sulfides, one should combine them all. But again a question can be raised: If one is interested in predicting the boiling point of *smaller* sulfides, why does one need information of compounds that are *twice* its size? It is a matter of philosophy, and while we appreciate the merits of studying a large data basis, we also appreciate the advantages of studying small homogeneous sets of compounds. Such a study focuses attention at different aspects of structural chemistry. In fact, one of the present author made numerous studies on the large set of compounds using diverse types of molecular descriptors.<sup>40–45</sup>

Concerning “overfitting”, which is clearly undesirable, we would like to point out that this is out of the question when one uses a single descriptor. Overfitting is a danger in multiple regression analysis when one uses too many

descriptors and has too few data. One cannot have overfitting with a single descriptor. This problem received some attention.<sup>46</sup> Does the variation of descriptors during the regression poses such a threat? Definitely so, just as a selection of descriptors from a large pool of descriptors (e.g. in CODESSA software) does the same. The difference between the two is that typically when using variable connectivity index, one generates about 40 different numerical alternative descriptors to choose from, CODESSA typically chooses a half-dozen descriptors from a pool of some 400 descriptors!

Finally we have to emphasize that while the idea of modifying chemical graph descriptors to differentiate heteroatoms is not new, as is well-illustrated by the pioneering work of Kier and Hall on valence connectivity indices,<sup>28</sup> the idea of modifying chemical graph descriptors to differentiate heteroatoms during the search for the best regression; that is, the idea of variable topological indices, is new.

#### ACKNOWLEDGMENT

This paper is contribution number 265 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by Grant F49620-94-1-0401.

#### REFERENCES AND NOTES

- Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; Chapter 10, pp 225–273.
- Balaban, A. T. Historical developments of topological indices. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds., in press.
- Randić, M. Topological Indices. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, U.K., 1998; pp 3018–3032.
- Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.
- Randić, M.; Novic, M.; Vracko, M. *Molecular Descriptors, New and Old*; Lecture Notes in Chemistry; Springer: Berlin, submitted for publication.
- Randić, M. Orthogonal molecular descriptors. *New J. Chem.* **1991**, *15*, 517–525.
- Randić, M. Resolution of ambiguities in structure–property studies by use of orthogonal descriptors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–370.
- Randić, M. Fitting of nonlinear regressions by orthogonalized power series. *J. Comput. Chem.* **1993**, *14*, 363–370.
- Randić, M. Curve fitting paradox. *Int. J. Quantum Chem, Quantum Biol. Symp.* **1994**, *21*, 215–225.
- Amić, D.; Davidović-Amić, D.; Jurić, A.; Lučić, B.; Trinajstić, N. Structure–activity correlation of flavone derivatives for inhibition of cAMP phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034–1038.
- Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. The structure–property models can be improved using the orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532–538.
- Šošćić, M.; Plavšić, D.; Trinajstić, N. Link between orthogonal and standard multiple linear regression models. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 829–832.
- Katritzky, A. R.; Lobanov, V.; Karelson, M. *CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis)*; University of Florida: Gainesville, FL, 1994.
- Basak, S. C. *POLLY*; (Natural Resources Research Institute, Duluth, University of Minnesota: Duluth, MN, 1988.
- Hall, L. H. *MOLCONN-X*; Hall Associates Consulting, Quincy: MA, 1991.
- Randić, M. On the characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
- Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
- Randić, M. Linear combinations of path numbers as molecular descriptors. *New J. Chem.* **1997**, *21*, 945–951.
- Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations 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.
- Randić, M. Novel graph theoretical approach to heteroatoms in QSAR. *Chemom. Intel. Lab. Syst.* **1991**, *10*, 213–227.
- Randić, M. On computation of optimal parameters for multivariate analysis of structure–property relationship. *J. Chem. Inf. Comput. Sci.* **1991**, *12*, 970–980.
- Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quantum Chem.* **1998**, *70*, 1209–1215.
- Randić, M.; Mills, D.; Basak, S. C.; Pogliani, L. On characterization of physical properties of amino acids. *New J. Chem.*, submitted for publication.
- Randić, M. High quality structure-property regressions. Boiling points of smaller alkanes. *New J. Chem.*, in press.
- Randić, M.; Basak, S. C.; Pompe, M.; Novic, M. Prediction of Gas Chromatographic Retention Indices Using Variable Connectivity Index. *Acta Chim. Slov.*, submitted for publication.
- Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. Molecular Connectivity I. Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971–1974.
- Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity Series Applied to Density. *J. Pharm. Sci.* **1975**, *65*, 1226–1230.
- Basak, S. C.; Magnuson, V. R. Determining structural similarity of chemicals using graph-theoretical indices. *Discrete Appl. Math.* **1988**, *19*, 17–44.
- Kier, L. B.; Hall, L. H. Molecular Connectivity VII. Specific Treatment of Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806–1809.
- Estrada, E. Edge adjacency relationships and novel topological index related to molecular volume. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 31–33.
- Randić, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261–266.
- Randić, M.; Basak, S. C. Multiple regression analysis with optimal molecular descriptors. *SAR QSAR Environ. Res.*, in press.
- Randić, M.; Pompe, M. On characterization of CC double bond in Alkenes. *SAR QSAR Environ. Res.* **1999**, *10*, 451–471.
- Randić, M.; Pompe, M. Work in progress.
- Timmermans, B. M. W. M.; van Zweiten, P. A. Quantitative structure–activity relationship in centrally acting imidazolidines structurally related to clonidine. *J. Med. Chem.* **1977**, *20*, 1636–1644.
- Based on the reviewers' comments.
- Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington D.C., 1995.
- Kubinyi, H. *Hansch Analysis and Related Approaches*; VCH: Weinheim, Germany, 1993.
- Basak, S. C.; Niemi, G. J.; Veith, G. D. Optimal characterization of structure for prediction of properties. *J. Math. Chem.* **1990**, *4*, 185–205.
- Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. Prediction of octanol/water partition coefficient ( $K_{OW}$ ) with algorithmically derived variables. *Environ. Toxicol. Chem.* **1991**, *10*, 893–900.
- Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: Hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651–655.
- Basak, S. C.; Grunwald, G. D.; Niemi, G. J. Use of graph-theoretic and geometrical molecular descriptors in structure–activity relationships. In *From Chemical Topology to Three-Dimensional Geometry*; Balaban, A. T., Ed.; Plenum Press: New York 1997.
- Basak, S. C.; Niemi, G. J.; Veith, G. D. Recent developments in the characterization of chemical structure using graph-theoretical indices. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 235–277.
- Basak, S. C.; Gute, B. D.; Ghatak, S. Prediction of complement–inhibitory activity of benzamides using topological and geometric parameters. *J. Chem. Inf. Comput. Sci.*, submitted for publication.
- Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238–1244.