
Synergistic Interactions among QSAR Descriptors

STEPHEN C. PETERANGELO, PAUL G. SEYBOLD

*Department of Chemistry, Wright State University, 3640 Colonel Glenn Hwy.,
Dayton, Ohio 45435-0001*

Received 23 February 2003; accepted 26 February 2003

DOI 10.1002/qua.10591

ABSTRACT: Quantitative structure–activity relationships (QSARs) and quantitative structure–property relationships (QSPRs) rely on regression equations containing numerical descriptors of molecular structure. In constructing these models, highly correlated descriptors are sometimes excluded from the regression equations. Although this exclusion seems reasonable, in fact it can lead investigators to overlook significant descriptor combinations, because the small differences between highly correlated descriptors sometimes encode important structural information. Furthermore, the multicollinearity that results from employing correlated descriptors is not as serious a problem as is often assumed. Described are several examples of cases in which pairs of highly correlated, poorly performing single-parameter descriptors yield highly significant structure–property regression equations. In effect, the descriptors act synergistically and yield regression equations that model the systems examined better than the sum of the individual components. A discussion of practical approaches to this problem is given. © 2003 Wiley Periodicals, Inc. *Int J Quantum Chem* 96: 1–9, 2004

Key words: QSAR; QSPR; regression equations; synergistic interactions

Introduction

A common practice in constructing multiple linear regression (MLR) equations for quantitative structure–activity relationships (QSARs) and quantitative structure–property relationships (QSPRs) is to eliminate highly correlated descriptors [1–4]. On its face, this practice appears quite reasonable. Highly correlated variables clearly con-

tain redundant information that might be more effectively encoded by a single variable. Furthermore, and most importantly from a statistical point of view, correlated independent variables lead to multicollinearity, which can cause problems in interpreting the results of a regression equation. Another, less common, practice is to focus attention on the best single descriptors in forming a regression equation.

Multicollinearity is a genuine concern in statistical analysis, but unfortunately one that often leads to much misunderstanding in the QSAR/QSPR

Correspondence to: P. G. Seybold; e-mail: paul.seybold@wright.edu

field. Perfect multicollinearity occurs when one of the independent variables in a regression equation is perfectly correlated with another variable or a linear combination of other variables [5]. In this situation, it is impossible to calculate least-squares estimates for the parameters. The situation of perfect multicollinearity is easily handled, and uncommon, but lesser degrees of multicollinearity are quite common, and their diagnosis and assessment can be an important part of the model-building process.

Contrary to common belief, the primary problem with high multicollinearity is not that the overall regression equation suffers. In fact, multicollinearity has no effect on the overall fitness of a model, so multicollinearity is irrelevant in a model whose purpose is strictly predictive (as opposed to explanatory) in nature [5–7]. The primary problem with multicollinearity is that it increases the standard errors associated with the individual regression coefficients, thereby decreasing their value for purposes of interpretability [5–7]. One method of diagnosing multicollinearity is to determine the correlation between the variable of interest and other variables in the regression equation, with values close to 1 implying high multicollinearity. The correlation values can then be used to calculate variance inflation factors (VIFs), defined as [6]

$$\text{VIF} = \frac{1}{(1 - R_j^2)} \quad (1)$$

where R_j^2 is the coefficient of determination between the j th coefficient regressed against all the other independent variables in the model. This is an arbitrary statistic in that the null hypothesis is not being accepted or rejected [6]. Because of this, there is no distinct cut-off value for a VIF. A common procedure is to set the value at 10, with any number higher indicating “serious” multicollinearity [6]. This practice is not entirely adequate, as it does not account for the overall fitness of the model under consideration. An alternative, proposed by Freund and Wilson [6], is to compare the VIF with the equivalent statistic for the entire model, a measure here called the model inflation factor (MIF):

$$\text{MIF} = \frac{1}{(1 - R_{\text{model}}^2)} \quad (2)$$

where R_{model}^2 is the coefficient of determination for the model. A VIF greater than the MIF would imply

a stronger relationship among independent variables than their relationship to the dependent variable.

To the best of our knowledge, Randić was the first to point out the problems inherent in the policy of focusing on the best descriptors and eliminating highly correlated descriptors [8–10]. Taking the molar refractions of the nine heptane isomers as a property example, he showed that although the molecular connectivity indices $^1\chi$ and $^2\chi$ are individually very poor single descriptors for this property ($r = 0.0241$ and $r = 0.1635$, respectively) and are also highly correlated ($r = 0.9810$), taken together the two descriptors yield a quite respectable regression for the property ($r = 0.9646$) [8]. Although in this particular example the single parameter $^3\chi$ yields an even better MLR account ($r = 0.9708$) for the molar refractions, the demonstration is remarkable. Kubinyi addressed much the same point, emphasizing that “... single X variables which are not correlated with the dependent variable must not be eliminated from the data set. Certain variables contribute only in combination with other variables” [11]. However, with some few exceptions [12, 13] these cautions appear to have been largely ignored, and the practice of focusing on “best” single descriptors and pruning highly correlated descriptors is frequently followed and is sometimes used as a default in commercial MLR software packages.

Here we use examples to show that this policy can cause QSAR and QSPR practitioners to overlook significant and perhaps meaningful correlations. We also show that two highly correlated descriptors may “interact” with a third descriptor very differently and yield quite different MLR results.

Methods

Data sets were taken from the literature as described. All the statistical studies were carried out using the software programs Microsoft Excel™ and QSARIS™ [14]. Normally, QSARIS employs a genetic algorithm to select MLR descriptors. For the purposes of the current analysis, however, all single descriptors yielding property regressions with $r > 0.65$ were ignored, and the remaining descriptors were examined pairwise both for their performances as MLR descriptors for various properties and for their mutual correlations.

TABLE I
Regression statistics for cases examined.

Case	Descriptor(s)	<i>n</i>	<i>R</i> ²	<i>s</i>	<i>F</i>	<i>q</i> ²	<i>t</i> -ratio	VIF	MIF
1	⁰ χ	71	0.0045	75.23	0.31	−.0467	0.56	n/a	n/a
	¹ κ _α	71	0.1286	70.39	10.18	0.0839	3.19	n/a	n/a
	⁰ χ, ¹ κ _α	71	0.9289	20.25	444.1	0.9242	29.73, −27.67	10.63	14.06
2	<i>I</i> _x	19 (17)	0.2324 (0.2384)	0.4755 (0.4997)	5.15 (4.70)	−0.0608 (−0.0900)	−2.27 (−2.17)	n/a	n/a
	<i>Q</i> _{sv}	19 (17)	0.0009 (0.0004)	0.5425 (0.5725)	0.01 (0.007)	−.5031 (−0.6124)	0.122 (0.084)	n/a	n/a
	<i>I</i> _x , <i>Q</i> _{sv}	19 (17)	0.9037 (0.9741)	0.1736 (0.0954)	75.1 (263.2)	0.8676 (0.9666)	−12.25, 10.56 (−22.94, 19.94)	3.55 (3.79)	10.37 (38.8)
3	² χ	161	0.3498	23.65	85.53	0.3269	9.25	n/a	n/a
	³ χ _{vc}	161	0.0003	29.33	0.054	−0.0228	0.232	n/a	n/a
	² χ, ³ χ _{vc}	161	0.8471	11.49	439.0	0.8406	29.63, 22.71	2.54	3.54
4	<i>H</i> _{other}	161	0.3633	23.41	90.74	0.3423	9.53	n/a	n/a
	<i>S</i> _{dssc}	161	0.0132	29.14	2.13	−0.0083	1.46	n/a	n/a
	<i>H</i> _{other} , <i>S</i> _{dssc}	161	0.9000	9.31	710.7	0.8959	37.42, 29.11	1.91	10.0

Following the technique suggested by Randić [8], “orthogonal complements” σ_i from each descriptor pair $\{x_1, x_2\}$ were obtained using the residuals of regressions of the form $x_i = ax_j + b$. In this way two orthogonal complement functions, σ_1 and σ_2 , were constructed for each pair of descriptors, the subscript referring to the dependent variable in the regression equation employed to obtain the residuals (σ_1 referring to the modeled descriptor with the larger coefficient in the original equation). As Randić noted, the two functions σ_1 and σ_2 can yield different regression results [8].

strongly correlated ($r = 0.9518$) with each other, so that one might well be tempted to eliminate both from further consideration. Nonetheless, when the two descriptors were used together, they gave a quite respectable ($r = 0.9609$) two-term MLR account of the BPs for these halocarbons:

$$\text{BP(K)} = 130.2(\pm 4.38)^1\kappa_\alpha - 153.7(\pm 5.56)^0\chi + 266.30(\pm 7.85) \quad (3)$$

$$n = 71 \quad R^2 = 0.9289 \quad s = 20.3 \text{ K}$$

$$F = 444 \quad q^2 = 0.9242$$

Results and Discussion

CASE 1: BOILING POINTS OF HALOGENATED HYDROCARBONS

The boiling points (BPs) of a set of 71 halogenated methanes, ethanes, and ethylenes compiled by Dixon and Seybold [15] were examined using molecular connectivity indices. As seen in Table I, both the zeroth-order connectivity index ⁰χ and Kier’s kappa-alpha shape index ¹κ_α [16] were poor single descriptors for the BPs, showing $r = 0.0810$ and $r = 0.3793$, respectively. Thus by itself, ⁰χ accounts for less than 1% of the variance in the BP and ¹κ_α, less than 15%. These two descriptors were also

Here n is the number of compounds, R^2 is the coefficient of determination, s is the standard error, F is the Fisher statistic for the regression, and q^2 is the leave-one-out (LOO) statistic. The numbers in parentheses refer to the standard errors of the coefficients. There are only two descriptors in this equation, so they have the same VIF when regressed against each other. In this case, VIF = 10.41, but the MIF = 13.05, so we can say that there is less correlation between the descriptors than their combination shows to the BP. Furthermore, as previously mentioned, the primary concern with multicollinearity is that uncertainty is introduced into the regression coefficients. The t -statistics for the two

descriptors are both above 25, and for 70 degrees of freedom, the critical value t must exceed for a P of 0.0005 is just 3.435 [6]. In this case, there is clearly no problem with multicollinearity affecting the reliability of the coefficients.

Based on the improvement above, it is of interest to examine the performance of the "orthogonal complements" σ_1 and σ_2 of the functions ${}^0\chi$ and ${}^1\kappa_\alpha$ in modeling the halocarbon BPs. In fact, σ_2 performs fairly well as a single descriptor:

$$\text{BP(K)} = -153.18(\pm 9.26)\sigma_1 + 307.87(\pm 3.98) \quad (4)$$

$$n = 71 \quad R^2 = 0.7961 \quad s = 33.81 \text{ K} \quad F = 273$$

This is illustrated in Figure 1(b). In contrast, in this case, σ_2 yields only a poor model of the BPs ($R^2 = 0.4511$, $s = 55.48$, $F = 58$).

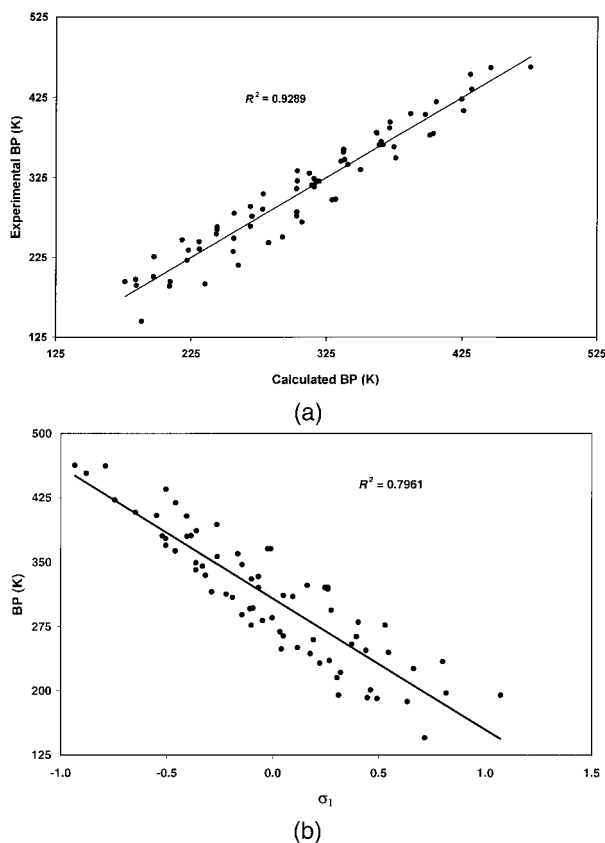


FIGURE 1. Plot of the experimental halocarbon boiling points versus (a) BPs calculated using a two-term regression with descriptors ${}^0\chi$ and ${}^1\kappa_\alpha$, Eq. (1); and (b) BPs calculated using the orthogonal complement function σ_1 .

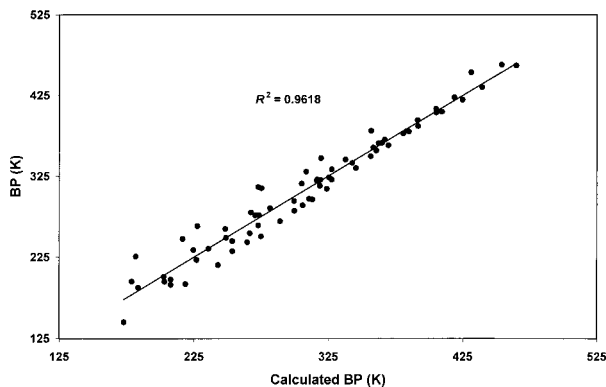


FIGURE 2. Plot of the experimental halocarbon BPs versus BPs calculated using the three-term regression of Eq. (3).

The descriptors ${}^0\chi$ and ${}^1\kappa_\alpha$ are both related to the extent of branching, although they account for branching in different ways. For example, ${}^1\kappa_\alpha$ takes account of the presence of heteroatoms, whereas ${}^0\chi$ does not. Thus, although these two descriptors cover much of the same topological territory, it is their difference that supplies crucial information for the BPs. We note that of the six compounds in the data set with the highest residuals in regression (2) above, all are methanes, which necessarily have no differences in branching, except to the extent that heteroatoms are involved.

The simple first-order connectivity index ${}^1\chi$, like ${}^0\chi$ and ${}^1\kappa_\alpha$, shows little correlation ($r = 0.1168$) with the halocarbon BPs, but it is strongly correlated with both ${}^0\chi$ ($r = 0.9801$) and ${}^1\kappa_\alpha$ ($r = 0.9380$). Nonetheless, addition of ${}^1\chi$ to the regression of Eq. (1) above does noticeably improve it:

$$\begin{aligned} \text{BP(K)} = & 128.1(\pm 3.25){}^1\kappa_\alpha - 198.5(\pm 7.19){}^0\chi \\ & + 109.1(\pm 14.3){}^1\chi + 244.14(\pm 8.05) \quad (5) \end{aligned}$$

$$n = 71 \quad R^2 = 0.9618 \quad s = 15.0 \text{ K}$$

$$F = 561.6 \quad q^2 = 0.9576$$

The calculated and predicted BPs for this example are plotted in Figure 2. In this case, there are three descriptors, so VIFs must be calculated for each by regressing them against the other two. For the coefficients in Eq. (5), the VIFs are ${}^1\kappa_\alpha = 10.71$, ${}^1\chi = 25.58$, ${}^0\chi = 32.70$. The MIF for this model is 31.35. This indicates that for this model, only the ${}^0\chi$ term shows an unacceptable degree of multicollinearity, and then only by a small amount. The t -statistics for

the coefficients are all well above the $P = 0.0005$ range, implying that multicollinearity may not, in fact, be a problem. Based on the contradictory results for results of multicollinearity, the choice is left to the discretion of the investigator.

Despite the high correlation noted above between ${}^1\chi$ and ${}^0\chi$, the combination of ${}^1\chi$ with ${}^1\kappa_\alpha$ yields a much poorer regression for the halocarbon BPs ($r = 0.7230$) than does ${}^0\chi$ and ${}^1\kappa_\alpha$ ($r = 0.9609$).

CASE 2: INHIBITION OF CYTOCHROME P-450 BY ALIPHATIC ALCOHOLS

Cohen and Mannering [17] examined the inhibition of cytochrome P-450 by a variety of aliphatic alcohols. The data were later analyzed (using weighted paths) by Amić et al. [18] in terms of $\log_{10}(1/C_{50})$, where C_{50} is the alcohol concentration (in mM) required for 50% inhibition. In the current case we consider two different descriptors: I_x , the moment of inertia about the x -axis, and Q_{sv} , a molecular polarity index comprising sums of intrinsic E -state values (see Ref. 20, p. 65–67). Taken by themselves, neither descriptor shows much correlation with $\log_{10}(1/C_{50})$ (I_x , $r = -0.4821$; and Q_{sv} , $r = 0.0296$), and they are strongly correlated with each other ($r = 0.8457$). Yet, taken together, these descriptors yield a good description of the data:

$$\log_{10}(1/C_{50}) = -0.024(\pm 0.0020)I_x + 0.651(\pm 0.616)Q_{sv} - 5.57(\pm 0.06) \quad (6)$$

$$n = 19 \quad R^2 = 0.9036 \quad s = 0.174 \\ F = 75.1 \quad q^2 = 0.8676$$

When two questionable data values are removed from the data set, the correlation improves to $R^2 = 0.9742$, with $s = 0.095$, and $F = 263$, with both t -statistics being above 19. The VIFs for the model are 3.55 (3.79 with $n = 17$). Compared with the MIF of 10.37 (38.8), these are insignificant. Again, the t -statistics are very robust, at -12.25 (-22.94) for I_x , and 10.56 (19.94) for Q_{sv} , indicating that multicollinearity is not a problem. The data excluding the two questionable points are plotted in Figure 3(a).

For this system, the orthogonal complement function σ_2 yields an excellent model:

$$\log_{10}(1/C_{50}) = -0.0243(\pm 0.0019) \\ \times [-0.0259(\pm 0.0012)]\sigma_2$$

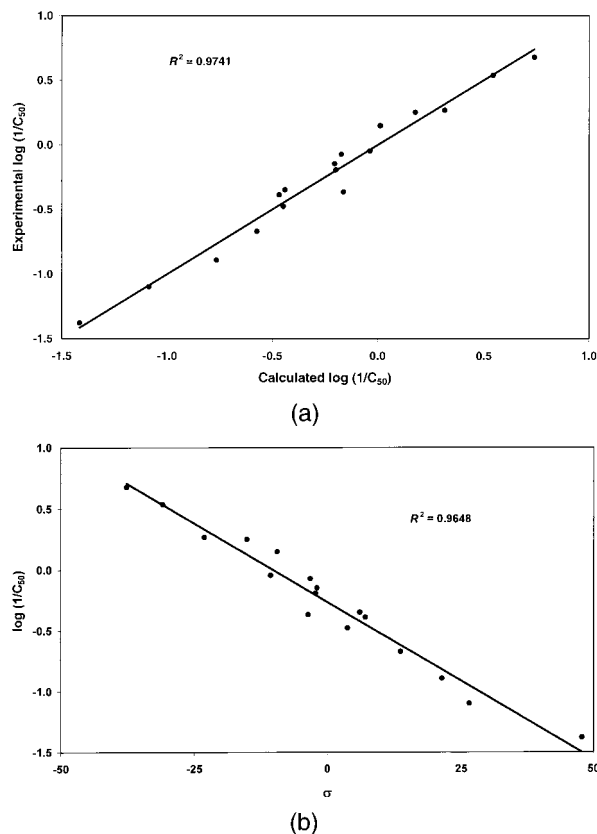


FIGURE 3. Plot of experimental cytochrome P-450 inhibition $\log_{10}(1/C_{50})$ versus (a) results determined using Eq. (4); and (b) results determined using the orthogonal complement function σ_1 .

$$-0.2699(\pm 0.0388) \\ \times [-0.2641(\pm 0.0261)] \quad (7)$$

$$n = 19 \quad (17) \quad R^2 = 0.9028 \quad (0.9648) \\ s = 0.1692 \quad (0.1075) \quad F = 158 \quad (411)$$

(Note that the figures in parentheses are for the case $n = 17$.) The results for $n = 17$ are plotted in Figure 3(b). The results for σ_2 show that the difference between these two descriptors carries significant explanatory information. In this case, σ_1 shows a poorer result ($R^2 = 0.6714$ [0.7604], $s = 0.311$ [0.2803], $F = 35$ [48]).

CASE 3: BOILING POINTS OF MONOALKENES

Recently, Nelson and Seybold [19] examined a number of physical properties of a large set of

monoalkenes. As a single descriptor for the BPs of these compounds, the simple second-order connectivity index ${}^2\chi$ shows only a modest correlation ($n = 161$, $r = 0.5914$) and the third-order valence cluster index ${}^3\chi_c^v$ almost no correlation ($r = 0.0173$). The intercorrelation of these two descriptors was $r = 0.7790$. When both descriptors are used together, however, one obtains an MLR equation with $r = 0.9204$:

$$\text{BP(K)} = 73.4(\pm 3.36){}^2\chi - 76.1(\pm 2.48){}^3\chi_c^v + 208.39(\pm 5.62) \quad (8)$$

$$n = 163 \quad R^2 = 0.8471 \quad s = 11.5 \text{ K} \\ F = 440 \quad q^2 = 0.8406$$

[We note that the standard error in the BPs for this set is smaller than that for the halocarbon data set because more than twice as many alkene BP values were available and because the alkenes boil over a much smaller range (157 K) than do the halocarbons examined in the first example (292 K).] Again, the VIFs are below the MIF (2.54 versus 3.54), and the t -statistics demonstrate that there is little doubt that the coefficients are statistically significant.

Here the orthogonal complement function σ_2 yields

$$\text{BP(K)} = 73.41(\pm 2.47)\sigma_2 + 46.72(\pm 10.95) \quad (9)$$

$$n = 161 \quad R^2 = 0.8468 \quad s = 11.45 \text{ K} \quad F = 884$$

These results are summarized in Table I and illustrated in Figure 4(a). The orthogonal function σ_1 is less successful in modeling the BPs result ($R^2 = 0.4964$, $s = 20.77$, $F = 158$).

CASE 4

For the same alkene data set as in the previous case, two other descriptors, H_{other} and S_{dssC} , can be examined. The electrotopological state descriptor H_{other} sums the electrotopological state values for nonpolar hydrogens (in the current case those in C–H bonds) [20]. It gives $r = 0.5925$ when measured against the 161 BPs. The electrotopological state sum index S_{dssC} represents double-bonded carbons with two single bonds to other carbons. It shows almost no correlation with the alkene BPs ($r = 0.0969$). The correlation between these two descrip-

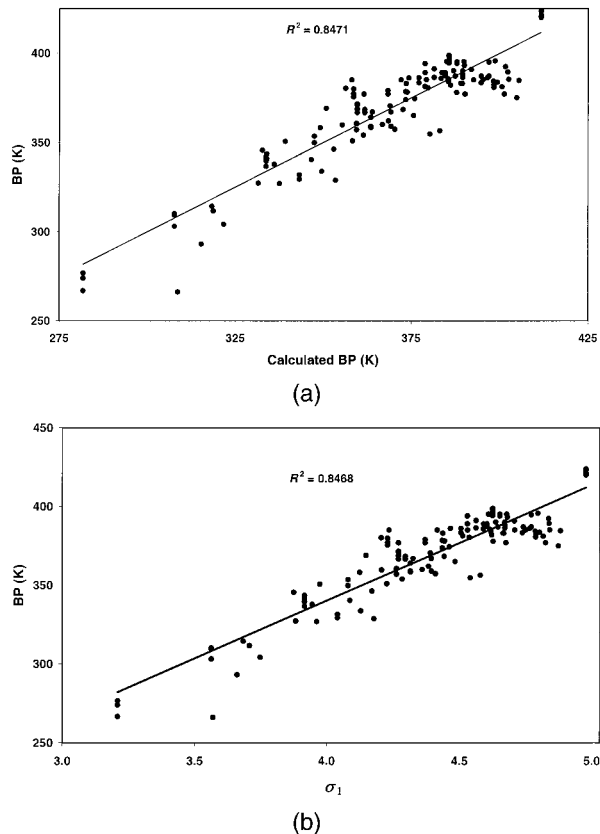


FIGURE 4. Plot of the experimental monoalkene boiling points versus (a) BPs calculated using the two-term regression in Eq. (6); and (b) BPs calculated using the orthogonal complement function σ_1 .

tors was $r = -0.6027$. Taken together, however, they yield a respectable account of the alkene BPs:

$$\text{BP(K)} = 57.65(\pm 1.54)H_{\text{other}} + 34.84(\pm 1.20)S_{\text{dssC}} + 144.56(\pm 6.02) \quad (10)$$

$$n = 161 \quad R^2 = 0.9000 \quad s = 9.31 \text{ K} \\ F = 710.7 \quad q^2 = 0.8959$$

This is illustrated in Figure 5(a). Again, although there is strong correlation between the descriptors, it is less than to their combined correlation to the response variable, and the t -tests show absolutely no problem with coefficient uncertainty.

The orthogonal complement σ_1 yields the following results

$$\text{BP(K)} = 57.61(\pm 1.64)\sigma_1 + 370.9(\pm 0.78) \quad (11)$$

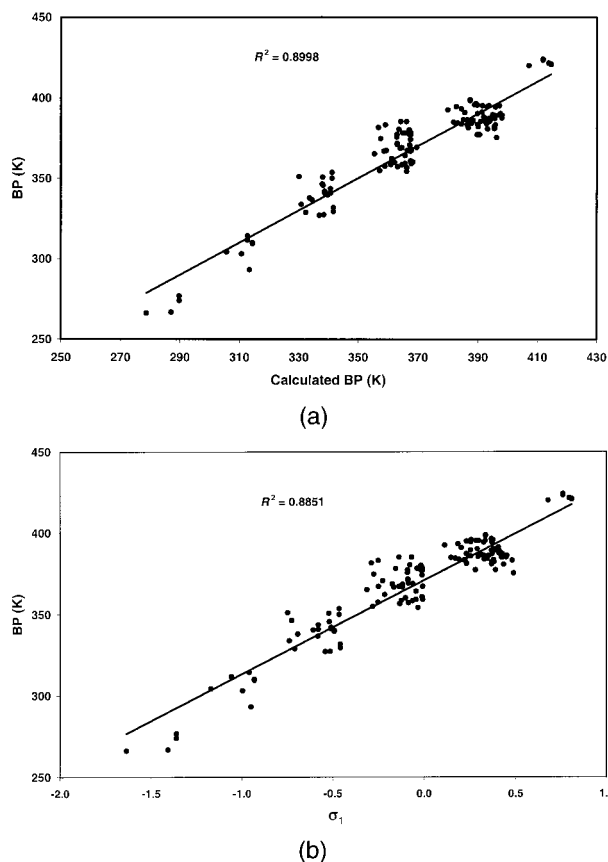


FIGURE 5. Plot of the experimental monoalkene BPs versus (a) BPs calculated using the two-term regression in Eq. (8); and (b) BPs calculated using the orthogonal complement function σ_1 .

TABLE II
Regression statistics for orthogonal complement functions.

Case	Descriptor	n	R^2	s	F
1	σ_1	71	0.7961	33.81	273
	σ_2	71	0.3835	55.48	58
2	σ_1	19 (17)	0.9028 (0.9648)	0.1691 (0.1075)	158 (411)
	σ_2	19 (17)	0.6714 (0.7604)	0.3111 (0.2803)	35 (48)
3	σ_1	161	0.8468	11.45	884
	σ_2	161	0.4964	20.77	158
4	σ_1	161	0.8851	9.92	1233
	σ_2	161	0.5487	19.66	195

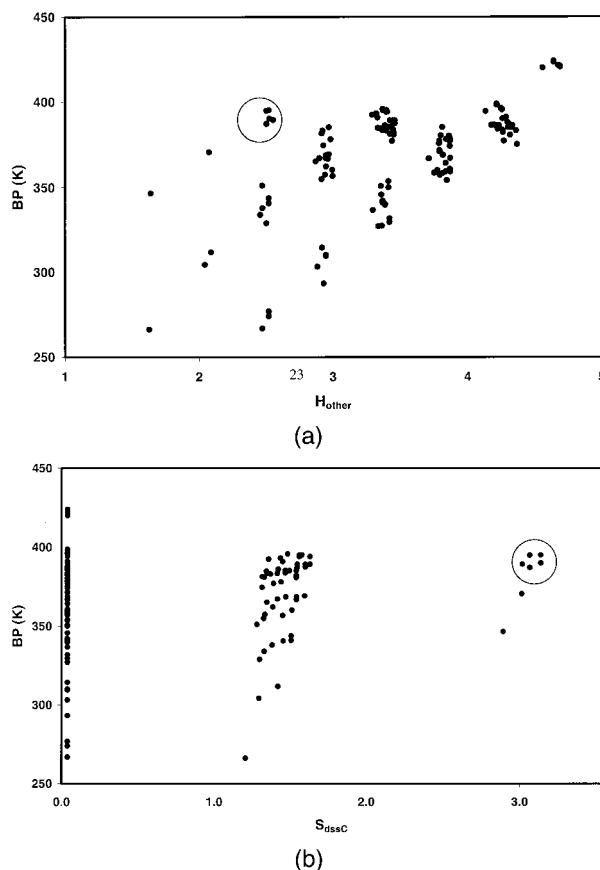


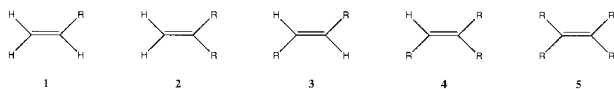
FIGURE 6. Plot of the experimental monoalkene BPs versus (a) BPs calculated using a single-term regression with descriptor H_{other} and (b) BPs calculated using a single-term regression with descriptor S_{dssC} .

$$n = 161 \quad R^2 = 0.8851 \quad s = 9.92 \text{ K} \quad F = 1232$$

These results are summarized in Table II and illustrated in Figure 5(b). The orthogonal function σ_2 is less successful in modeling the BPs result ($R^2 = 0.5487$, $s = 19.66$, $F = 194$).

This particular example contains a further point of interest. The individual descriptors H_{other} and S_{dssC} , as a result of their definitions, show a certain degree of clustering, that is, they fall into clumps of similar values for the monoalkenes. Single-variable regression plots of the BPs versus the two individual parameters, as shown in Figure 6, illustrate this. This clustering arises because there are (excluding ethylene, which is unique) only a limited number of topological arrangements possible about the double bond, and in the nature of the electrotopological parameters the parameters from each arrangement show only limited variations in their values. The

possible arrangements can be summarized as follows:



The descriptor S_{dssC} sums the electrotopological state values for the carbons in the double bond, and these fall generally into three classes according to the number of carbons involved in the double bond that are connected to two other carbons. Types **1** and **3** both have no double bonded carbons connected to two other carbons, types **2** and **4** each have one, and type **5** has two. Types **1** and **2** represent exterior double bonds, unbranched and branched, respectively, whereas types **3–5** represent interior double bonds with increasing degrees of branching about the double bond, with type **5** being uncommon. The circles in Figure 6 enclose compounds from type **5**. The descriptor H_{other} has a similar but complementary relationship to the BPs, so that when combined, the two descriptors yield a good description of the BPs, whereas individually they do not.

Conclusions

The examples given here demonstrate that Randić's observation [8] that highly correlated, poorly performing, single descriptors can nonetheless supply important descriptive information was not an isolated instance. It is important to recognize that the small differences between such descriptors—the so-called “orthogonal complements” of the descriptors—can in some cases provide useful structural input in building MLR equations and relations. Thus, the common practice of discarding one of the correlated pair of descriptors from further consideration is not warranted. The results clearly support Randić's contention that “the criteria for inclusion or exclusion of descriptors should not be based on parallelism between descriptors even if overwhelming . . .” [21]. It should also be realized that although two descriptors may be highly correlated, multicollinearity may not necessarily be a problem. Furthermore, even if high multicollinearity does exist, it is not a problem if the model is to be used only for predictive purposes.

What, then, can one do to improve the efficiency of forming MLR expressions? One very useful and informative approach was suggested by Randić himself in his “orthogonal descriptors” technique

[8–10, 21]. In this approach, a number of different descriptor sets and orders are examined, and, having chosen a starting descriptor, subsequent descriptors are added only as their orthogonal complements to the descriptors already present. This approach is essentially an application of the well-known Gram–Schmidt orthogonalization procedure [22]. It has the advantages that the regression coefficients are stable (i.e., do not change as new descriptors are added), and the new information supplied by each additional descriptor is clearly distinguished in the regression statistics. Because the order in which the descriptors are introduced is significant, it is important in this approach to examine different orders for entry. Experience has made it clear that the natural inclination of starting with the best single descriptor does not always lead to the best final multiple-descriptor model [8, 12].

An alternative is to rely on a genetic algorithm procedure [11] to pick the best set of descriptors. However, in this approach it is important to refrain from employing any default procedure that removes pairs (or multiples) of strongly correlated descriptors. Such preliminary pruning can eliminate potentially important structural information lurking in the orthogonal complements of such descriptors.

ACKNOWLEDGMENT

One of the authors (P.G.S.) thanks Prof. Milan Randić for helpful discussions related to this topic and for calling attention to Ref. 13.

References

1. Unger, S. H.; Hansch, C. *J Med Chem* 1973, 16, 745–749.
2. Jurs, P. C. *Anal Chem* 1981, 53, 2184–2187.
3. Jurs, P. C.; Yuta, K. *J Med Chem* 1981, 24, 241–251.
4. Godden, J. W.; Xue, L.; Bajorath, J. *J Chem Inf Comput Sci* 2002, 42, 1263–1269.
5. Lewis-Beck, M. *Applied Regression: An Introduction*; Sage: Beverly Hills, 1986; 58–64.
6. Freund, R.; Wilson, W. *Regression Analysis Statistical Modeling of a Response Variable*; Academic: London, 1998; 181–223.
7. Berry, W. D.; Feldman, S. *Multiple Regression in Practice*; Sage: Beverly Hills, 1985; 37–50.
8. Randić, M. *New J Chem* 1991, 15, 517–525.
9. Randić, M. *J Chem Inf Comput Sci* 1991, 31, 311–320.
10. Randić, M.; Seybold, P. G. *SAR and QSAR in Environ Res* 1993, 1, 77–85.
11. Kubinyi, H. *Quant Struct-Act Relat* 1994, 13, 393–401.

12. Lučić, B.; Nicolčić, S.; Trinajstić, N.; Juretić, D. *J Chem Inf Comput Sci* 1995, 35, 532–538.
13. Xu, L.; Zhang, W.-J. Comparison of Different Methods for Variable Selection. *Analyt Chim Acta* 2001, 446, 477–483.
14. SciVision, 200 Wheeler Rd., Burlington, MA 01803.
15. Dixon, S.; Seybold, P. G. (to be published).
16. Kier, L. B.; Hall, L. H. In Boyd, D.; Lipkowitz, K., Eds. *Reviews in Computational Chemistry*, Vol. 2; VCH Publishers: New York, 1991; Chap. 9, p. 367–422.
17. Cohen, G. M.; Mannering, G. J. *Molec Pharm* 1973, 9, 383–397.
18. Amić, D.; Lucić, B.; Nikolić, S.; Trinajstić, N. *Croatia Chem Acta* 2001, 74, 237–250.
19. Nelson, S. D.; Seybold, P. G. *J Molec Graphics Modelling* 2001, 20, 36–53.
20. Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*. Academic Press: New York, 1999.
21. Randić, M. J. *Chem Inf Comput Sci* 1997, 37, 672–684.
22. Kaplan, W. *Advanced Calculus*. Addison-Wesley: Reading, 1952; p. 164–165.