High quality structure–property regressions. Boiling points of smaller alkanes

Milan Randić

Department of Mathematics and Computer Science, Drake University, Des Moines, IA 50311, USA. FAX: +1 515 292 8629; e-mail: milan.randic@drake.edu

Received (in Gainsville, FL, USA) 16th June 1999, Revised manuscript received 19th January 2000, Accepted 21st January 2000

For the correlation of the boiling points of smaller alkanes we outline the use of the variable connectivity index $^1\chi^f$ involving a parameter x, to be determined during the regression fitting. The optimal values of the variable x is determined by minimizing the standard error of the regression. With the value $x = 0.65$ we obtained a quadratic regression characterized by a standard error of 2 °C, when out of 21 structures one structure is eliminated as an outlier (having a standard error over 6 °C). This result, based on a single descriptor, is better than hitherto reported regressions on the same set of compounds using other descriptors. Using as an additional descriptor the product P_3W_3, where P_3 are the paths of length three and W_3 are the walks of length three, the standard error was further reduced to 1.56 °C. We include also the regression results derived for orthogonalized descriptors and clarify the alternative approaches to orthogonalization of descriptors.

It has been recognized for some time that the success or the failure of regression analysis critically depends of the selection of appropriate molecular descriptors. This in a way may have been responsible for the proliferation of molecular descriptors (graph theoretical, geometrical, or quantum chemical). For example, the statistical package CODESSA[1] evaluates hundreds of various molecular descriptors and selects typically half a dozen as the best for the structure–property or structure–activity relation considered. In screening combinatorial libraries one may focus attention on 20–30 molecular descriptors (out of hundreds), which are used to filter out a dozen or two dozen most promising candidate structures out of 100 000 or a million. For example, recently Lahana and coworkers[2] were able to select 26 peptides from some 280 000 structures using topological indices and other molecular descriptors, which were further analyzed and eventually led to the synthesis of a compound that has 100 times higher activity than the initial compound. In Table 1 we illustrate the topological indices employed by Lahana and coworkers, indicating the range of acceptable values for the indices considered.

The selection of descriptors for multivariate regression analysis (MRA), as well as for combinatorial screening analysis (CSA), could be improved significantly if we could design descriptors that not only have a better discriminatory power but would produce results of a similar quality based on fewer descriptors. Recently the problem of selection of descriptors from a large pool of descriptors has received due attention.[3–6] In selection of the descriptors researchers often exclude an index that is highly interrelated with one of the indices already selected.[7,8] This is not only unnecessary, having no theoretical justification, but can also lead to inferior results! Consider, as an illustration, the regression of the molar refraction (MR) for octane isomers[9,10] based on the first order connectivity index[11] $^1\chi$ and the second order connectivity index[12] $^2\chi$. The two connectivity indices have a very high mutual correlation: the coefficient of regression r is equal 0.9757. According to the prevailing practice if we have selected $^1\chi$ as a descriptor for a regression, then we would have to exclude $^2\chi$ as a descriptor that does not introduce sufficient novelty. With $^1\chi$ in simple regression for MR we obtain a regression with the coefficient $r = 0.087$. If we use $^2\chi$, we obtain again a regression with a low correlation coefficient ($r = 0.177$), as could be expected because of the mutual interrelation. If, however, we use both $^1\chi$ and $^2\chi$ in a two-descriptor regression, we obtain a regression with a surprisingly high correlation coefficient of $r = 0.971$. Hence, the two-parameter regression was successful not because $^1\chi$ and $^2\chi$ parallel each other but because the small part in which $^2\chi$ differs from $^1\chi$ is relevant for the property considered.

The above example clearly points to the direction that is essential for the selection of relevant molecular descriptors, which is the *complementarity* of descriptors, not their *parallelism*. This is true whether we use two, three, or twenty descriptors, but clearly one would prefer to use as few descriptors as possible.

Table 1 Topological indices and their ranges used in screening a combinatorial library

Descriptor	Minimum	Maximum
Kappa alpha 2	26.1	44.3
Flexibility	22.5	40.3
Kier Chi v4	3.325	5.342
Balaban index	2.846	6.701
Kappa 1	56.1	93.0
Kappa 2	29.5	51.2
Kappa 3	22.1	41.8
Kappa alpha 1	51.8	86.9
Kappa alpha 3	19.3	37.2
Randic index	54.2	87.6
Wiener index	86 872	312 008
E-state	160.5	268.0

The connectivity index $^1\chi$

In 1975 Randić[11] introduced a bond additive connectivity index $^1\chi$ as a descriptor to characterize molecular branching. Kier and Hall[7,8] recognized the potential of the connectivity index not only for characterization of structure–property relationships but also for use in QSAR (quantitative structure–activity relationships). During the 25 years since the introduction of the connectivity index, we have witnessed a con-

Table 2 Modified adjacency matrix of 2,3-dimethylpentane and corresponding row sums

	1	2	3	4	5	6	7	Row sum
1	x	1	0	0	0	0	0	$1+x$
2	1	x	1	0	0	1	0	$3+x$
3	0	1	x	1	0	0	1	$3+x$
4	0	0	1	x	1	0	0	$2+x$
5	0	0	0	1	x	0	0	$1+x$
6	0	1	0	0	0	x	0	$1+x$
7	0	0	1	0	0	0	x	$1+x$

Bond	Connectivity index	Variable connectivity index
1–2	$1/\sqrt{(1\cdot 3)}$	$1/\sqrt{(1+x)(3+x)}$
2–3	$1/\sqrt{(3\cdot 3)}$	$1/\sqrt{(3+x)(3+x)} = 1/(3+x)$
3–4	$1/\sqrt{(2\cdot 3)}$	$1/\sqrt{(2+x)(3+x)}$
4–5	$1/\sqrt{(1\cdot 2)}$	$1/\sqrt{(1+x)(2+x)}$
2–7	$1/\sqrt{(1\cdot 3)}$	$1/\sqrt{(1+x)(3+x)}$
3–8	$1/\sqrt{(1\cdot 3)}$	$1/\sqrt{(1+x)(3+x)}$
Molecule	$3/\sqrt{3} + 1/\sqrt{2} + 1/3 + 1/6$	$3/\sqrt{(1+x)(3+x)} + 1/\sqrt{(1+x)(2+x)} + 1/(3+x) + 1/\sqrt{(2+x)(3+x)}$

siderable rise in the number of proposed molecular descriptors.[13–15] Nevertheless, the connectivity index $^1\chi$, the associated higher order connectivity indices[12] $^m\chi$, the valence connectivity indices[16,17] $^1\chi^v$, and their generalizations ("path-cluster" indices of Kier and Hall[18,19]) continue to be among the mostly widely used molecular descriptors even today.

The seminal paper on the connectivity index[11] has approached 1500 citations and continues to be cited. One may wonder why the connectivity index $^1\chi$ has been so successful, particularly for bond additive properties. The answer lies in its construction: the connectivity index differentiates bond types through weighting of bond contributions. It assigns to bonds that are "exposed," such as the terminal C–CH_3, C–CH_2, or C–CH bonds and the bridging CH_2–CH_2 bonds, greater weights than to the "internal" bonds between tertiary and quaternary carbon atoms. The weights used were selected because they represent a solution to a set of inequalities that reproduce the *ordering* of isomers of smaller alkanes that *parallels* the relative magnitudes of their boiling points.

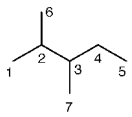

Fig. 1 Molecular skeleton and numbering of carbon atoms for 2,3-dimethylpentane.

Variable connectivity index $^1\chi^f$

In this contribution we want to draw attention to a novel generalization of the connectivity indices that puts them in a class by themselves and sets them apart from other topological indices. Topological indices are characterized by fixed numerical values, which are independent of the property considered. Hence, they can be computed once the bonding pattern (or the geometry in the case of 3D structural indices) of a molecule is known. In contrast, the variable connectivity indices $^1\chi^f$ depend on the property considered. They were proposed some time ago[20,21] as a novel approach for the characterization of heteroatoms in chemical structures. In contrast to the approaches that assign different parameters to different heteroatoms,[13,14,22–25] we have a variable parameter that will undergo change *during* the regression analysis.

The variable connectivity index $^1\chi^f$ offers a powerful tool for the study of physico-chemical molecular properties. Several recent publications demonstrated the use of the variable connectivity index $^1\chi^f$ on selected properties of

Table 3 Variable connectivity index $^1\chi^f$ for the smaller alkanes

	Alkane	Bond contributions
1	Ethane	$1/(1+x)$
2	Propane	$2/[(1+x)(2+x)]$
3	Butane	$2/\sqrt{[(1+x)(2+x)]} + 1/(2+x)$
4	Isobutane	$3/\sqrt{[(1+x)(3+x)]}$
5	Pentane	$2/\sqrt{[(1+x)(2+x)]} + 2/(2+x)$
6	Isopentane	$1/\sqrt{[(1+x)(2+x)]} + 2/\sqrt{[(1+x)(3+x)]} + 1/\sqrt{[(2+x)(3+x)]}$
7	Neopentane	$4/\sqrt{[(1+x)(4+x)]}$
8	Hexane	$2/\sqrt{[(1+x)(2+x)]} + 3/(2+x)$
9	2-M-pentane	$2/\sqrt{[(1+x)(3+x)]} + 2/\sqrt{[(2+x)(3+x)]} + 1/(2+x) + 1/\sqrt{[(1+x)(2+x)]}$
10	3-M-pentane	$2/\sqrt{[(1+x)(2+x)]} + 2/\sqrt{[(2+x)(3+x)]} + 1/\sqrt{[(1+x)(3+x)]}$
11	2,2-MM-butane	$3/\sqrt{[(1+x)(4+x)]} + 1/\sqrt{[(1+x)(2+x)]} + 1/\sqrt{[(2+x)(4+x)]}$
12	2,3-MM-butane	$4/\sqrt{[(1+x)(3+x)]} + 1/(3+x)$
13	Heptane	$2/\sqrt{[(1+x)(2+x)]} + 4/(2+x)$
14	2-M-hexane	$2/\sqrt{[(1+x)(3+x)]} + 2/(2+x)\ 1/\sqrt{[(2+x)(3+x)]} + 1/\sqrt{[(1+x)(2+x)]}$
15	3-M-hexane	$2/\sqrt{[(1+x)(2+x)]} + 1/(2+x) + 2/\sqrt{[(2+x)(3+x)]} + 1/\sqrt{[(1+x)(3+x)]}$
16	3-E-pentane	$3/\sqrt{[(1+x)(2+x)]} + 3/\sqrt{[(2+x)(3+x)]}$
17	2,2-MM-pentane	$3/\sqrt{[(1+x)(4+x)]} + 1/\sqrt{[(2+x)(4+x)]} + 1/(2+x) + 1/\sqrt{[(1+x)(2+x)]}$
18	2,3-MM-pentane	$3/\sqrt{[(1+x)(3+x)]} + 1/\sqrt{[(1+x)(2+x)]} + 1/\sqrt{[(2+x)(3+x)]} + 1/(3+x)$
19	2,4-MM-pentane	$4/\sqrt{[(1+x)(3+x)]} + 2/\sqrt{[(2+x)(3+x)]}$
20	3,3-MM-pentane	$2/\sqrt{[(1+x)(2+x)]} + 2/\sqrt{[(2+x)(4+x)]} + 2/\sqrt{[(1+x)(4+x)]}$
21	2,2,3-MMM-butane	$3/\sqrt{[(1+x)(4+x)]} + 2/\sqrt{[(1+x)(3+x)]} + 1/\sqrt{[(3+x)(4+x)]}$

Table 4 Numerical values of the variable connectivity index $^1\chi^f$ for selected values of x

	$x = -0.5$	$x = 0$	$x = 0.5$	$x = 0.65$	$x = 1$
1	2.0000	1.00000	0.66667	0.60606	0.50000
2	2.30940	1.41421	1.03280	0.95646	0.81650
3	2.97607	1.91421	1.43280	1.33381	1.14983
4	2.68328	1.73205	1.30931	1.22245	1.06066
5	3.64273	2.41421	1.83280	1.71117	1.48316
6	3.45995	2.27006	1.72733	1.61473	1.40403
7	3.02327	2.00000	1.53960	1.44408	1.26491
8	4.30940	2.91421	2.23280	2.08853	1.81650
9	4.12662	2.77006	2.12733	1.99209	1.73736
10	4.23662	2.80806	2.14535	2.00701	1.74740
11	3.85892	2.56066	1.96924	1.84616	1.61513
12	3.97771	2.64273	2.03146	1.90391	1.66421
13	4.97607	3.41421	2.63280	2.46589	2.14983
14	4.79329	3.27006	2.52733	2.36945	2.07070
15	4.90329	3.30806	2.54535	2.38437	2.08073
16	5.01329	3.34607	2.56338	2.39929	2.09077
17	4.52559	3.06066	2.36924	2.22352	1.94846
18	4.75438	3.18074	2.44948	2.29619	2.00758
19	4.61050	3.12590	2.42187	2.27301	1.99156
20	4.69413	3.12132	2.39888	2.24824	1.96535
21	4.39470	2.94338	2.27955	2.14076	1.87940

alcohols,[21,26] amines,[27] sulfides[28] and amino acids.[29] All of the mentioned studies involved molecules having heteroatoms. We want to show in this paper that the variable connectivity index can improve also the regressions of saturated hydrocarbons, molecules having all atoms and all bonds of the same kind. As we will see the primary, secondary, tertiary and quaternary carbon atoms affected differently the variable descriptor. This additional flexibility gives to the variable connectivity index the ability to better describe hydrocarbons.

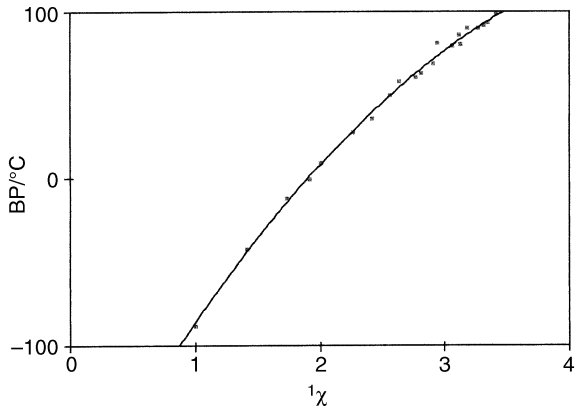

Fig. 2 Regression of the boiling points of smaller alkanes against the connectivity index $^1\chi$.

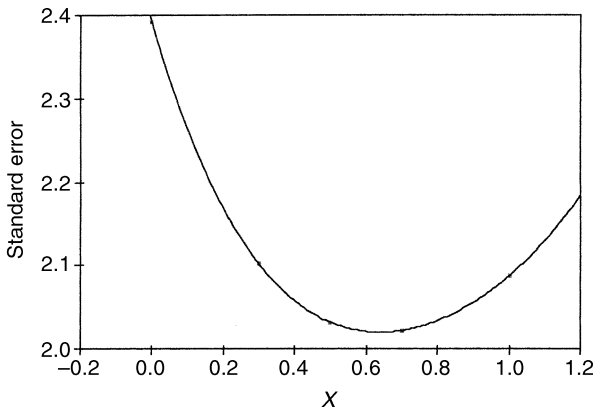

Fig. 3 Variation of the standard error of estimate with change of the diagonal entry x of the adjacency matrices.

Construction and properties of $^1\chi^f$

The connectivity index for molecules can be calculated from the adjacency matrix by summing the bond contributions given by $1/\sqrt{(S_i S_j)}$. Here S_i, S_j are the row sums corresponding to the bond (i, j) or to the row i and the row j of the adjacency matrix. The same is true for construction of the variable connectivity index, the difference being in the modification of the row sums. In Table 2 we show the modified adjacency matrix of 2,3-dimethylpentane (depicted in Fig. 1) in which, instead of the zeros on the main diagonal, we have a variable parameter x. Again the bond contributions are given by $1/\sqrt{(S_i S_j)}$ but now the row sums involve the variable x. In the lower part of Table 2 we illustrate calculation of the bond contributions to $^1\chi$ and $^1\chi^f$. In the case of saturated hydrocarbons the diagonal entry x modifies the relative weights of the carbon valences. Instead of the factors 1, $1/\sqrt{2}$, $1/\sqrt{3}$ and 1/2 that enter into the construction of the connectivity indices for primary, secondary, tertiary and quaternary carbon atoms we have $1/\sqrt{(1+x)}$, $1/\sqrt{(2+x)}$, $1/\sqrt{(3+x)}$ and $1/\sqrt{(4+x)}$, for CH_3, CH_2, CH and C, respectively. Clearly variation of x induces different perturbations for carbon atoms having different valences.

In Table 3 we list the variable connectivity index $^1\chi^f$ for alkanes having from two to seven carbon atoms. The labels M and E stand for methyl and ethyl groups, respectively, label MM for the dimethyl group, *etc.* If we set $x = 0$ the variable connectivity index $^1\chi^f$ reduces to the simple connectivity index $^1\chi$. However, for $x \neq 0$ the variable connectivity index is a function of the variable weight x.

In order to illustrate numerically the variation of $^1\chi^f$ with x we list in Table 4 the variable connectivity indices for the 21 smaller alkanes of Table 3 for five different values of x, including the case $x = 0$. The value $x = +0.65$, as will be seen later, corresponds to the optimal choice of x for the correlation of the boiling points of smaller alkanes. A close look at Table 4 reveals that for negative values of x the magnitude of the variable index increases, while for positive values the magnitudes, relative to the values of the connectivity index when $x = 0$, decrease. More important than the variations of the absolute magnitudes of the connectivity indices are the relative changes of the magnitudes for selected pairs of structures. For example, 2,4-dimethylpentane and 3,3-dimethylpentane (entries 19 and 20) have very close values for the simple connectivity index $^1\chi$, the former being slightly bigger. For positive values of x the difference between them becomes even more pronounced, while for negative values there is a reversal of their relative magnitudes. It is this flexibility of the variable connectivity index that allows it to better adjust for a given property regression.

Boiling points in smaller alkanes

To demonstrate the performance of the variable connectivity index $^1\chi^f$ we revisited the correlations of the boiling points in alkanes having from two to seven carbon atoms. These alkanes include bond types between primary, secondary, tertiary and quaternary carbons, except for the CC bond between two quaternary carbons, which appears among smaller alkanes for the first time in the octane isomer 2,2,3,3-tetramethylbutane. The same set of smaller alkanes has been used by Kier and Hall[7] to test the relative power of different descriptors.

That the assumption of linearity is inadequate for correlation between the connectivity index and BP of alkanes can be seen from Fig. 2, in which we plotted the boiling points of smaller alkanes against the connectivity index $^1\chi$. A visual

Table 5 The experimental boiling points, the calculated boiling points and residuals for optimal $^1\chi^f$ ($x = 0.65$) with and without 2,2,3-trimethylbutane (entry 21), and cross-validation for the regression obtained by excluding the outlier

	BP expt	BP calcd	Residual	BP* calcd	Residual	Cross-validation	
1	−88.63	−88.05	−0.58	−87.75	−0.88	−85.31	−3.32
2	−42.07	−43.49	+1.42	−43.47	+1.40	−43.87	+1.80
3	−0.50	−0.41	−0.09	−0.61	+0.11	−0.62	+0.12
4	−11.73	−12.59	+0.86	−12.73	+1.00	−12.90	+1.17
5	36.07	37.59	−1.52	37.24	−1.17	37.40	−1.33
6	27.85	28.36	−0.51	28.05	−0.20	28.08	−0.23
7	9.5	11.22	−1.72	10.97	−1.47	11.20	−1.70
8	68.74	70.49	−1.75	70.08	−1.34	70.18	−1.44
9	60.27	62.57	−2.30	62.16	−1.89	62.32	−2.05
10	63.28	63.82	−0.54	63.41	−0.13	63.42	−0.14
11	49.74	49.94	−0.20	49.57	+0.17	49.54	+0.14
12	57.99	55.03	+2.96	54.64	+3.35	54.31	+3.68
13	98.42	98.31	+0.11	97.91	+0.51	97.77	+0.65
14	90.05	91.68	−1.63	91.27	−1.22	91.45	−1.40
15	91.85	92.73	−0.88	92.32	−0.47	92.39	−0.54
16	93.48	93.77	−0.29	93.36	+0.12	93.34	+0.14
17	79.20	81.03	−1.83	80.61	−1.41	80.72	−1.52
18	89.78	86.43	+3.35	86.01	+3.77	85.62	+4.16
19	80.50	84.73	−4.23	84.31	−3.81	84.67	−4.17
20	86.03	82.89	+3.14	82.47	+3.56	82.16	+3.87
21	80.88	74.65	+6.23	*	*	*	*

inspection of Fig. 2 shows that the connectivity index, as a single descriptor, appears satisfactory. But can we do better? Can we further reduced the standard error of estimate of the boiling points of smaller alkanes with a *single* descriptor?

Optimal connectivity index for the boiling points of smaller alkanes

In Fig. 3 we show the variation of the standard error of the regression of **BP** for smaller alkanes as x varies from 0 (zero) towards +1. Each time we select a value for x to be tested we first have to calculate the corresponding connectivity indices $^1\chi^f$, and then examine the statistical parameters for a quadratic regression. We found a minimum in the curve around $x = +0.65$, which is the optimal value for the parameter x in this study. In Table 5 we list the experimental boiling points and the calculated boiling points with the corresponding residuals for the variable connectivity index (with $x = +0.65$). We find the standard error to be just below 2.5 °C. In comparison the simple connectivity index ($x = 0$) gives a standard error just below 3 °C. A drop of 0.5 °C for the standard error may appear small but it represents a significant improvement because we already had a relatively good regression when using $^1\chi$.

The regression eqn. (1) and (2), given in Table 6, correspond to correlations using the simple connectivity index and the variable connectivity index, respectively. That the variable connectivity index produced a significantly better regression is better seen by comparing the corresponding Fisher ratios, which from just below 3000 went above 4000 after optimization of the variable diagonal entry x of the modified adjacency matrices. A closer look at Table 5 shows that the residual for 2,2,3-trimethylbutane is well above the value of twice the standard error ($2s = 4.96$ °C), making this compound an outlier. If we discard 2,2,3-trimethylbutane from the correlation an additional drop of 0.5 °C occurs for the standard error, and the Fisher ratio exceeds 6000.

In the middle of Table 5 we give the computed boiling points when the outlier 2,2,3-trimethylbutane has been discarded. Observe that now all the residuals are well below 4 °C, the new limit for 2*s*. In order to verify the stability of the obtained regression in the last column of Table 5 we give the computed **BP** and the residuals for cross-validation (one-leave-out). As we can see from Table 5 by comparing the corresponding columns, the values of the new residuals are very close to the old ones.

Further improvement of a relatively satisfactory regression may appear to some as unwarranted. In our view minor improvements in regressions, particularly when arrived at by using fewer descriptors, are important for the following reasons. (1) Regressions of *high quality* with fewer descriptors may pinpoint dominant structural components responsible for the correlation more clearly then regressions using a larger number of descriptors in MRA. The role of individual descriptors is obscured by the presence of others, because of their interrelation. (2) Regressions of *high quality* may point to outliers, which (if not due to experimental inaccuracies) may suggest limitations of the descriptors used, and point to the direction in which further improvements of the regressions are possible. We will see both these aspects of the refinement of MRA by continuing to examine more closely the regression of the boiling points of smaller alkanes.

Regressions using two descriptors

The calculated **BP** for 2,2,3-trimethylbutane (74.65 °C) is too small. Having a single molecule with a tertiary-quaternary CC bond type does not justify changing the weighting algorithm for this bond type. So, we have to consider additional descriptors that could account for the crowded group of methyls occurring in 2,2,3-trimethylbutane. Wiener,[30] already in 1947, used paths of length three (P_3) as a descriptor that supple-

Table 6 The quadratic regression eqn. (1)–(3) using $^1\chi^f$ as descriptor

	Coeff.	Std. error	*t*-stat.	
Descriptors: $^1\chi^f$ ($x = 0$); $n = 21$				
x	133.5361	6.7058	19.9137	eqn. (1)
x^2	−12.9653	1.4242	−9.1037	
Cons.	−207.1542	7.4498	−27.8068	
	$r = 0.9985$	$s = 2.928$	$F = 2922$	
Descriptors: $^1\chi^f$ ($x = +0.65$); $n = 21$				
x	155.0934	6.7567	22.9540	eqn. (2)
x^2	−17.8684	2.0366	−8.7738	
Cons.	−175.4828	5.2592	−33.3668	
	$r = 0.9989$	$s = 2.481$	$F = 4073$	
Descriptors: $^1\chi^f$ ($x = +0.65$); $n = 20$				
x	153.8851	5.5133	27.9118	eqn. (3)
x^2	−17.5981	1.6600	−10.6012	
Cons.	−174.5515	4.2912	−40.6770	
	$r = 0.9993$	$s = 2.020$	$F = 6017$	

mented W (the Wiener number) and made "corrections" for crowded portions in highly branched alkanes.

When we include P_3 as an additional descriptor we obtain the regression eqn. (4) (Table 7) with a standard deviation of 1.71 °C. If we use W_3, the count of walks of length three, we do not obtain much improvement over the quadratic regression. However, if we use as a descriptor the product P_3W_3 we obtain regression eqn. (5) (Table 7) with a standard error of 1.57 °C. The plot of the calculated **BP** against the experimental **BP** is shown in Fig. 4. It is interesting to observe that now 2,2,3-trimethylbutane is no longer an outlier. Its residual (+1.36 °C) happens to be even smaller than the standard error! This clearly points out that 2,2,3-trimethylbutane was an outliner not because its **BP** was inaccurate but because the molecular descriptor used was deficient.

The connectivity indices, being bond additive, clearly have limitations when considering the effects caused by atoms and atomic groups beyond the nearest neighbors. Thus, just as they are not able to account for crowded methyl groups, they cannot account for differences in *meta-para* substitutions of the benzene ring. It is therefore not surprising that they fail in such applications,[31] in fact they should not even have been considered in such situations. This kind of misuse and misunderstanding of the connectivity indices may have provoked unwarranted skepticism towards methematical descriptors, and towards connectvtiviy indices in particular.[32] Of course, one should not expect that a single descriptor can adequately characterize all the diverse structural features of molecules. However, as has been demonstrated, the P_3W_3 descriptor can successfully characterize effects of close methyl groups occurring in alkanes.

If we eliminate two apparent outliers we obtain the regression equation shown as step 3 in Table 8, with associated standard error of 1.01 °C. This is the best result hitherto known for the regression on smaller alkanes when two descriptors are used. Observe that the range of the boiling points is almost 190 °C and that isomeric variations cover about 10 °C in the case of hexane isomers and almost 20 °C in the case of heptane isomers. Despite such large variations in the boiling points of smaller alkanes the two descriptors $^1\chi^f$ and P_3W_3 were able to well characterize the size and the shapes of these compounds.

Table 7 The quadratic regression eqn. (4) and (5) using $^1\chi^f$, $(^1\chi^f)^2$ and P_3 or P_3W_3 as descriptors

	Coeff.	Std. error	t-stat.	
Descriptors $^1\chi^f$ ($x = +0.65$), P_3; $n = 21$				
x	154.7974	4.6508	33.2841	eqn. (4)
x^2	−19.8253	1.4653	−13.5303	
P_3	1.8740	0.4089	4.5831	
Cons.	−173.4755	3.6461	−47.5785	
	$r = 0.9995$	$s = 1.708$	$F = 5739$	
Descriptors $^1\chi^f$ ($x = +0.65$), P_3W_3; $n = 21$				
x	157.5588	4.2919	36.7111	eqn. (5)
x^2	−20.1499	1.3560	−14.8597	
P_3W_3	0.0280	0.0053	5.3054	
Cons.	−175.5129	3.3210	−52.8491	
	$r = 0.9996$	$s = 1.567$	$F = 6819$	

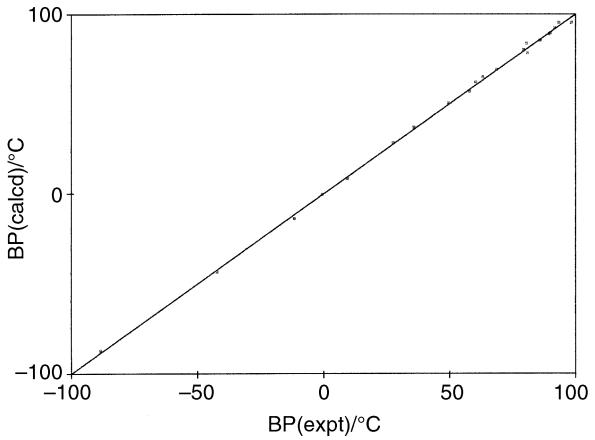

Fig. 4 Plot of boiling points using the description P_3W_3 calculated against the experimental boiling points.

Orthogonal regression equation

Multivariate regression analysis (MRA), as is well known, has been plagued by the instability of the regression equations. Each time a new descriptor is included in a stepwise regression (Table 8), the coefficients of the regression equation prior to the introduction of a new descriptor change; often they change dramatically, displaying the inherent inconsistency of the stepwise regression equations. This malady that has typified MRA since its introduction, and which was until relatively recently tacitly ignored by practitioners, has been corrected as outlined elsewhere.[33–36]

The instability of regression equations has been traced to the interrelation of the descriptors. If the descriptors used are only weakly interrelated the changes of the coefficients in the regression equation upon introduction of an additional descriptor are minor. However, when the descriptors used are highly interrelated, which often is the case, the changes in regression equations are large. A way out of these difficulties, that accompany traditional MRA, is to use molecular descriptors that are not related. Since there are no such thing, one has to make them not unrelated by modifying the descriptors. Interrelated descriptors can be transformed into orthogonal (*i.e.*, unrelated) descriptors by following an orthogonalization procedure. The procedure is general and it can be applied regardless of how much or how little the descriptors used are interrelated. The orthogonality can also be extended to non-linear regression[35] and to similarity/dissimilarity studies.[37]

One starts the orthogonalization with a simple regression using a single descriptor as illustrated in step 1 in Table 9. Because at this stage this is the only descriptor it can be assumed orthogonal and be given the label $^1\Omega$ or alternatively Res 1/0 (which will become clearer later). The Greek letter omega has been deliberately selected to label the orthogonal descriptors as a reminder that this is the ultimate form for

Table 8 Stepwise regression equations using $^1\chi^f$, $(^1\chi^f)^2$ and P_3W_3 as descriptors. The bottom of the table shows the correlation matrix

	Coeff.	Std. error	t-stat.	
Step 1				
x	97.9755	2.4553	39.9039	
Cons.	−135.0852	4.6612	−28.9806	
	$r = 0.9947$	$s = 5.369$	$F = 1592$	
Step 2				
x	154.1321	6.7970	22.6763	
x^2	−17.4528	2.0850	−8.3705	
Cons.	−175.0041	5.1995	−33.6578	
	$r = 0.9990$	$s = 2.386$	$F = 4066$	
Step 3				
x	161.0477	2.9971	53.7344	
x^2	−21.6047	1.0090	−21.4122	
P_3W_3	0.0334	0.0039	8.5858	
Cons.	−175.5129	3.3210	−52.8491	
	$r = 0.9998$	$s = 1.014$	$F = 15051$	

	BP	x	x^2	P_3W_3
BP	1.000	0.995	0.967	0.808
x		1.000	0.987	0.818
x^2			1.000	0.852
P_3W_3				1.000

Table 9 Stepwise regression equations using Res 1/0, Res 2/1, and Res 3/2 as orthogonal descriptors. The bottom of the table shows the correlation matrix

	Coeff.	Std. error	*t*-stat.	
Step 1				
Res 1/0	97.9755	2.4553	39.9039	
Cons.	−135.0852	4.6612	−28.9806	
	$r = 0.9947$	$s = 5.369$	$F = 1592$	
Step 2				
Res 1/0	97.9753	1.0914	89.7716	
Res 2/1	−17.4523	2.0854	−8.3690	
Cons.	−135.0850	2.0719	−65.1974	
	$r = 0.9990$	$s = 2.386$	$F = 4066$	
Step 3				
Res 1/0	97.9752	0.4635	211.3886	
Res 2/1	−17.4523	0.8856	−19.7067	
Res 3/2	0.0334	0.0039	8.5859	
Cons.	−135.0849	0.8799	−153.5229	
	$r = 0.9998$	$s = 1.014$	$F = 15051$	
	BP	Res 1/0	Res 2/1	Res 3/2
BP	1.000	0.995	−0.093	0.040
Res 1/0		1.000	0.000	0.000
Res 2/1			1.000	0.000
Res 3/1				1.000

molecular descriptors. In the next step, before introducing the second descriptor in the regression, one correlates the second descriptor $({}^1\chi^f)^2$ against the first descriptor ${}^1\chi^f$. The residuals of this regression are by definition those parts of the second descriptor, $({}^1\chi^f)^2$, that do not correlate with the first descriptor ${}^1\chi^f$. Hence, our second orthogonal descriptor is the residual between the second and the first descriptor, Res 2/1 or alternatively ${}^2\Omega$. The third orthogonal descriptor is obtained by first correlating P_3W_3 against ${}^1\chi^f$, which gives the residual Res 3/1. This residual, while orthogonal to the first descriptor Res 1/0, will not be orthogonal to the second orthogonal descriptor Res 2/1. To make it orthogonal to the second descriptor we consider the regression of Res 3/1 against Res 2/1. The residual of this regression, Res 3/2, is our third orthogonal descriptor ${}^3\Omega$. In the case of additional descriptors the outlined procedure is continued until all the descriptors are made orthogonal.

The lower part of Table 9 gives the orthogonalized regression equation with two and three descriptors. We give the statistical parameters *r*, *s* and *F* merely to indicate that they do not change upon orthogonalization (*cf.* the corresponding values in Tables 8 and 9). However, observe that all the standard errors for the coefficients of the regression equations decrease in comparison with the corresponding uncertainties of the coefficients of non-orthogonal regressions, which increase at each successive step in a stepwise regression. This is in accordance with previous observations concerning orthogonalized molecular descriptors.[36] As a consequence, when orthogonalized descriptors are used the *t*-statistic improves at each step, while the opposite is true with the use of non-orthogonal descriptors.

The distinction between orthogonal and non-orthogonal descriptors is also reflected in the correlation matrix, which has been included in Tables 8 and 9 for the non-orthogonal and the orthogonal descriptors, respectively. The entries in the correlation matrices show the degree of pairwise interrelation. In the case of non-orthogonal descriptors (Table 8, the bottom part) we can see that the descriptors are very interrelated. The first row of the correlation matrix similarly indicates that each of the descriptors selected has an appreciable correlation with **BP**. In the case of an orthogonalized descriptor the submatrix belonging to the interrelations among the descriptors is the identity matrix, confirming that indeed the descriptors are not related. The first row of the correlation matrix again tells us how much each of the descriptors correlates with **BP**. Since the first descriptor already accounts for 99% of the variance clearly the remaining descriptors have but a minor (but important) contribution. The squares of the entries in the first row of the correlation matrix of the orthogonalized descriptors add up to 1. Hence, from the correlation matrix we can immediately deduce the percentage that each descriptor accounts for the correlation of the property. Therefore, for ${}^1\chi^f$, the orthogonal part of $({}^1\chi^f)^2$, and the orthogonalized part of P_3W_3, we obtain as the percentages 99%, 0.85% and 0.15%, respectively. The corresponding sum of squares of the correlation entries in the first row of the correlation matrix for non-orthogonal descriptors is meaningless.

Finally we should add, in order to avoid confusion, that there are alternative approaches to the orthogonalization of molecular descriptors. Our orthogonalization procedure, although based on regressions among residuals, can be alternatively viewed also as an illustration of the Gram–Schmidt orthogonalization of vectors. We do not use the inner product (scalar product) in order to extract the projections among descriptors. Instead, we use *residuals* of regressions between the descriptors considered as a way to extract the orthogonal parts of descriptors. However, one can interpret the orthogonalization of descriptors (vectors) as outlined by Lučić and Trinajstić,[3] which has an analogous expression in quantum chemistry,[38] as a Gram–Schmidt orthogonalization, if one normalizes descriptors.[39] An illustration of the standard Gram–Schmidt orthogonalization of connectivity indices based on Galois theory was recently given by Araujo and Morales.[40]

Comparison with alternative approaches

In order to emphasize the high quality of the regression obtained using the flexible connectivity index we will review a few results based on alternative molecular descriptors. Balaban and coworkers,[41] by applying information-theory formulas to a sequence of numbers obtained at increasing distance from the center of a molecule, arrived at a novel topological index IBC. The index IBC in a quadratic regression and combined with *N*, the number of carbon atoms as a second descriptor, leads for the **BP** of smaller alkanes (including octanes) a regression with $r = 0.9865$, the root mean square error $s = 6.78\,^\circ$C, and the Fisher ratio $F = 835$. If we now confine the analysis only to alkanes from C_3 to C_7, which makes the comparison with our correlation more appropriate, one obtains similar results: $r = 0.9895$, $s = 6.31$ and $F = 250$. These statistical parameters should be compared with the statistical parameters shown in Table 6 where we used a *single* descriptor in the quadratic regression. The simple connectivity index gave: $r = 0.9985$; $s = 2.93$ and F about 3000; while the variable connectivity index gave: $r = 0.9989$; $s = 2.48$; F just above 4000.

In another study Bonchev and coworkers[42] constructed novel 3D molecular descriptors by viewing molecules as systems of mutually repulsing atoms connected by covalent bonds of constant length. From the so-optimized geometry they extracted weights as a metric analogue of the vertex distance sum in molecular graphs. These 3D weights were then used to replace vertex degrees as used in several well-known topological (2D) indices, like the connectivity indices ${}^m\chi$ and the Zagreb group indices M_k.[14] They considered smaller alkanes including nonanes, so we recalculated their regression by limiting alkanes to the C_3–C_7 size range, to make it the same as our set (except that we also included in our regressions ethane, C_2, which could only adversely affect our results). Their index ${}^3D^0\chi$, when used as a *single* descriptor gave: $r = 0.9877$, $s = 6.44$, and $F = 717$. This result they compared to: $r = 0.9914$, $s = 6.75$, and $F = 1085$, which is obtained when ${}^1\chi$ is used as a *single* descriptor in a *linear*

regression. However, a *quadratic* regression using $^1\chi$ as descriptor gives much better results: $r = 0.9966$, $s = 2.93$, and $F = 2922$. If, however, $3D^0\chi$ is used in a quadratic regression their regression statistic is not improved at all ($r = 0.9877$, $s = 6.62$, and $F = 340$).

Using *three* descriptors Bonchev *et al.*[42] obtained much better results:

Descriptors: $3D^0\chi$, DSRW2, M_2 give:

$$r = 0.9925,\ s = 5.33,\ \text{and}\ F = 325$$

Descriptors: $3D^0\chi$, $^0\chi$, M_2 give:

$$r = 0.9988,\ s = 2.16,\ \text{and}\ F = 2174$$

Descriptors: $3D^0\chi$, 3D M_2, M_2 give:

$$r = 0.9991,\ s = 1.88,\ \text{and}\ F = 2856$$

The above results should be compared to our results shown in Table 7 (where we use quadratic regression and *two* descriptors). Again, using fewer descriptors, we obtain better regressions, with a standard error of 1.71 and 1.57 °C, as compared with the standard error of 1.88 °C of Bonchev and coworkers.

Concluding remarks

Use of "flexible" molecular descriptors can significantly improve regression analysis. This is true when one uses paths of variable weight,[43,44] or a variable connectivity index $^1\chi^f$. As we have seen in this study $^1\chi^f$ has decreased the standard error of a quadratic regression for the boiling points of smaller alkanes from 2.93 °C, corresponding to a quadratic regression using the ordinary ("fixed") connectivity index $^1\chi$, to 2.48 °C for $^1\chi^f$. Upon removal of an outlier the standard error has further dropped to the respectable value of 2 °C. Even greater improvements are possible when additional descriptors are used. Use of P_3 reduced the standard error to 1.71 °C and if the product of paths and walks of length three are used, P_3W_3, the standard error dropped to 1.57 °C.

Each time the standard error is reduced a possibility exists of the emergence of new outliers. The standard error of 1.57 °C has produced two outliers, *n*-heptane and 2,4-dimethylpentane. Upon their elimination we obtain a regression, using as descriptors $^1\chi^f$, $(^1\chi^f)^2$ and P_3W_3, with an impressive standard error of only 1.01 °C. Incidentally, this paper is the first occasion that P_3W_3 has been used as a descriptor. Paths as molecular descriptors have already been mentioned by Platt almost 50 years ago,[45] and P_3 has already been employed in the historic pioneering work of Wiener.[30] Walks[46–48] and weighted walks[49] have also been suggested as molecular descriptors. Very recently the quotient P_k/W_k has been suggested[50] as an alternative shape index to the kappa shape indices of Kier.[51–53]

Regressions with a standard error of about 1 °C can be viewed as a "high quality regressions." One of the purposes for trying to increase the accuracy of the regression equations is to identify the specific structural features that make some molecules outliers. Close examination of molecules showing large residuals may point to structural components that particular molecular descriptors fail to properly account for. This then will help in designs of better descriptors and will lead to even better regressions. The additional descriptor P_3W_3 used here was arrived at in an *ad hoc* manner, hence, we may expect that a systematic search for the second descriptor will reduce even further the already impressive standard error reported here.

References

1 A. R. Katritzky, V. S. Lobanov and M. Karelson, *Chem. Soc. Rev.*, 1995, **24**, 279.
2 G. Grassy, B. Calas, A. Yasri, R. Lahana, J. Woo, S. Iyer, M. Kaczorek, R. Floc'h and R. Buelow, *Nature Biotech.*, 1998, **16**, 748.
3 B. Lučić and N. Trinajstić, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 121.
4 B. Lučić, N. Trinajstić, S. Sild, M. Karelson and A. R. Katritzky, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 610.
5 A. R. Katritzky, E. S. Ignachenko, R. A. Barcock and V. S. Lobanov, *Anal. Chem.*, 1994, **66**, 1799.
6 M. Randić and M. Pompe, *J. Chem. Inf. Comput. Sci.*, to be submitted.
7 L. B. Kier and L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
8 L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Ltd., Letchworth, England, 1986.
9 M. Randić, *Croat. Chem. Acta*, 1993, **66**, 289.
10 M. Randić, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 672.
11 M. Randić, *J. Am. Chem. Soc.*, 1975, **97**, 6609.
12 L. B. Kier, W. J. Murray, M. Randić and L. H. Hall, *J. Pharm. Sci.*, 1976, **65**, 1226.
13 M. Randić, in *Topological Indices*, in *Encyclopedia of Computational Chemistry*, ed. P. Rague von Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III and P. R. Schreiner, John Wiley and Sons, Chichester, 1998, pp. 3018–3032.
14 N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, FL, 2nd edn., 1992, ch. 10, pp. 225–273.
15 R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley, New York, in the press.
16 L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, 1976, **65**, 1806.
17 L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, 1981, **70**, 583.
18 L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, 1983, **72**, 1170.
19 L. B. Kier and L. H. Hall, *J. Pharm. Sci.*, 1978, **67**, 1743.
20 M. Randić, *Chemometrics Intell. Lab. Syst.*, 1991, **10**, 213.
21 M. Randić, *J. Comput. Chem.*, 1991, **12**, 970.
22 E. J. Kupchik, *Quant. Struct.-Act. Relat.*, 1989, **8**, 98.
23 A. T. Balaban, *MATCH*, 1986, **21**, 115.
24 O. Ivancius, T. Ivanciuc and A. T. Balaban, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 395.
25 L. Pogliani, *J. Phys. Chem.*, 1999, **103**, 1598.
26 M. Randić, S. C. Basak, M. Novič and M. Pomple, *Acta Chim. Slov.*
27 M. Randić and J. C. Dobrowolski, *Int. J. Quant. Chem.*, 1998, **70**, 1209.
28 M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, submitted.
29 M. Randić, D. Mills, S. C. Basak and L. Pogliani, *New J. Chem.*, submitted.
30 H. Wiener, *J. Am. Chem. Soc.*, 1947, **69**, 17.
31 R. L. de Compadre, C. M. Compadre, R. Casillo and W. J. Dunn, III, *Eur. J. Med. Chem.*, 1983, **18**, 569.
32 H. Kubinyi, *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, Germany, 1993, p. 50 and references therein.
33 M. Randić, *New J. Chem.*, 1991, **15**, 517.
34 M. Randić, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 311.
35 M. Randić, *J. Comput. Chem.*, 1993, **14**, 363.
36 M. Randić, *Int. J. Quant. Chem: Quant. Biol. Symp.*, 1994, **21**, 215.
37 M. Randić, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1092.
38 A. Szabo and N. Ostlund, *Modern Quantum Chemistry*, McGraw-Hill, New York, 1989, pp. 15–21.
39 M. Randić, *J. Chem. Inf. Comput. Sci.*, submitted.
40 O. Araujo and D. A. Morales, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 1031.
41 A. T. Balaban, S. Bertelsen and S. C. Basak, *MATCH*, 1994, **30**, 55.
42 A. Toropov, A. Toropova, T. Ismailov and D. Bonchev, *J. Mol. Struct. (THEOCHEM)*, 1998, **424**, 237.
43 M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 261.
44 M. Randić and S. C. Basak, *SAR QSAR Environ. Res.*, in press.
45 J. R. Platt, *J. Phys. Chem.*, 1952, **56**, 328.
46 M. Randić, *J. Comput. Chem.*, 1980, **1**, 386.
47 Y. Jiang, *Sci. Sinica, Ser. B*, 1984, **27**, 236.
48 G. Rücker and C. Rücker, *J. Chem. Inf. Comput. Sci.*, 1991, **31**, 422.
49 D. Bonchev, X. Liu and D. J. Klein, *Croat. Chem. Acta*, 1993, **66**, 141.
50 M. Randić, *Chemometrics Intell. Lab. Syst.*, submitted.
51 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1985, **4**, 109.
52 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1986, **5**, 1.
53 L. B. Kier, *Quant. Struct.-Act. Relat.*, 1986, **5**, 12.

Paper b000780n