Pergamon

Bioorganic & Medicinal Chemistry (2002) 3981–3996

BIOORGANIC & MEDICINAL CHEMISTRY

On the Topological Evidences for Modelling Lipophilicity

Vijay K. Agrawal,[a] Jyoti Singh[a] and Padmakar V. Khadikar[b,*]

[a]*QSAR and Chemical Laboratories, A.P.S. University, Rewa-486 003, India*
[b]*Research Division, Laxmi Fumigation and Pest Control Pvt. Ltd. 3 Khatipura, Indore-452 007, India*

Received 5 April 2002; accepted 6 July 2002

Abstract—Topological evidences for modelling lipophilicity of a large series of diversed compounds have been provided on the basis of distance-based topological indices. A pool of topological indices along with indicator parameters related to the type of the compounds present in the set of 140 compounds were used for this purpose. The results have shown that topology as well as the type of compounds are the responsible parameters for modelling lipophilicity.

Introduction

The partitioning of drugs (i.e., organic compounds acting as drugs) between aqueous and lipophilic phase is of utmost importance for drug potency. Drug membrane interaction, drug transport, distribution, accumulation, protein and receptor binding, efficacy and drug resistance are influenced by drug lipophilicity.[1,2] Lipophilicity is an important endpoint used extensively in medicinal chemistry and environmental toxicity in predicting biological and hazardous effects of chemicals. Therefore, the development of new methods for the quantitative estimation of the lipophilicity of organic molecules is extremely important. The useful measure of relative lipophilicity of chemicals is their partition coefficient between *n*-octanol and water. In some of the recent studies non-empirically based molecular similarity methods in assessing physico-chemical and toxic properties of different groups are discovered.[3–5] No other physico-chemical property has attracted as much interest in quantitative structure–activity relationship studies (QSAR) as lipophilicity (synonymously called hydrophobicity), any differentiation between both terms is only a sematic nemicity; the opposite of lipophilicity is hydrophobicity.

New methods for the quantitative estimation of the lipophilicity of organic compounds is extremely important. These methods can be classified: (i) statistical calculation of increment values corresponding to occurring molecular substructures and (ii) correlation of logP data with molecular properties such as atomic charges, hydrogen bond effect, molecular volume, surface area and so on. The latter type of approach is feasible only if quantitatively characterized calculation procedures based on the additive constitutive character of logP have also been developed.[6,7] However, hydrophobic fragmental constants are not available for every structural characteristic, so logP predictions are not always possible.

Due to its predominant role in biological activity and despite the large number of papers already published, lipophilicity still constitutes a challenging subject for further investigation. Although a great deal is known about the experimental determination and the calculation of logP, the most widely accepted measure of lipophilicity, the problem of investigating the logP values of new compounds is far from being solved in many cases. Consequently, recently Schaper et al.[7] proposed that artificial neural networks (ANNs) are able to recognize structural features influencing lipophilicity and to directly use the chemical structure for the calculation of logP. Similarly, very recently we have proposed novel estimation of lipophilicity in that we have used newly introduced PI (Padmakar-Ivan) index[8,9] Also, we have modelled lipophilicity using information-theoretic topological index[10] Id. The primary aim of the investigation was to provide topological (structural) evidences for modelling logP and not to provide new methods for the calculation of logP. Another goal of the present investigation is to study relative potential of the newly

*Corresponding author. Tel.: +91-731-531906; fax: +91-7662-42175; e-mail: vijay-agrawal@lycos.com

0968-0896/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S0968-0896(02)00299-7

introduced Szeged index[11–15] (Sz). The characterization and some of the chemical applications of Sz index are reported in the literature.[16–25] However, to date no attempt has been made to investigate the potential of Sz index in modelling lipophilicity/hydrophobicity (logP). The results as discussed below establish that among the pool of distance-based topological indices, Sz is a better index for modelling lipophilicity.

The basic which determines the topological conditioning of the reactivity is the principle of 'molecular structure', according to which, molecules are considered as isolated objects, possessing a relatively rigid and permanent location of nuclei (atoms), joined each other by electronic forces (chemical bonds) which are highly specific and strongly localized. Hence, molecules are assumed to have a structure which conditions their physical and chemical properties. Therefore, as a consequence of this principle, it is hardly surprising that the most of 'natural' topological descriptors (distance-based topological indices) we have employed renders such satisfactory results. It is worth mentioning that although current topological indices can be compared in a nearly simple way, they are chosen on the grounds that in spite of their high collinearity they have different information contents. The approach has been applied to a set of 140 molecules, including various structural types (Table 1). The topological indices used for investigating relative potential of Sz index for modelling logP of these set of 140 compounds being: Wiener index[26] (W), first-order (χ) connectivity index[27] Branching index[28–30] (B), Balaban index[31,32] (J), hyper-Wiener index[33] (W*), and log RB.[28] The logP values needed for the study were adopted from the 'star list' of Hansch et al.[3]

The paper is organized along the following lines: next section deals with the most significant results discussing the relative merits of the Sz index. We then close the work discussing the values of the results and stating the main correlations of these findings and pointing out some possible future extension of the present methodology. Then we give some basic definitions of the topological indices, methodology used, and several necessary associated antecedents on this issue.

Results and Discussion

Based on the experimental part mentioned below, we now discuss the results related to relative potential of Sz index for modelling lipophilicity/hydrophobicity (logP) for a set of 140 compounds under present study (Table 1). The various structural types present in this set of 140 compounds include: simple mono-functional derivatives like hydrocarbons, alcohols, halides, amines, carboxylic acids, esters, ketones, phenols, furans, pyrroles, thiophenes, and pyridines, as well as some representatives of somewhat complex molecules with poly-functional groups.

The set of 140 compounds, their lipophilicity (logP) and indicator parameters used are given in Table 1. The calculated values of the distance based topological indices (W, Sz, B, χ, J, W* and logRB) are given in Table 2. Their inter-relations and correlations with logP are presented in Table 3 (correlation matrix). The estimated regression coefficients and their qualities are presented in Table 4. The obtained statistics is summarized in Table 5. The predictive potential of the proposed models is discussed on the basis of cross-validation and other parameters presented in Table 6. The correlation of experimental and calculated logP using most appropriate models are demonstrated in Figures 1–3. Finally, diagnostics to test for the presence of compounds outliers and multi-collinearity between descriptors in the model a variable inflation factor (VIF) is determined and summarized in Table 6.

A perusal of Tables 1 and 2 show that the magnitude of the topological indices used (W, B, J, Sz, $^1\chi^v$, logRB and W*) goes on increasing with increase in the bulk of the compounds used. All these topological indices exhibit low to high degeneracy. This is obvious as all of them (except $^1\chi^v$) belong to first-generation topological indices[32,34] According to Balaban[32,34] first-generation topological indices in spite of their degeneracy are very successful in modelling structure–property as well as structure–activity relationships (QSPRs, QSARs). As discussed below this is found to be the case in present study also.

Initial statistical analysis[35–37] has indicated that no statistically significant mono-parametric regression expressions (models) are possible for modelling lipophilicity (logP) of the compounds (140) used, and that fruitful correlations are obtained only in multi-parametric regressions, that too, containing 9, 10 and 11 molecular descriptors. Hence, we have to attempt only multi-parametric regressions with the combinations of 9, 10 and 11 descriptors which will provide us most appropriate models for modelling lipophilicity (logP) of the compounds used (Table 1).

Before a multi-parametric (multivariate) analysis is undertaken, it is convenient to tailor the data in certain ways to make the calculations easier. Normally it is sufficient to pre-process the data by means of auto-scaling and mean centering. Auto-scaling gives each variable unit variance and hence the same chance to contribute to a calculated model, whereas mean centering facilitates interpretation. In doing so, the first step in developing appropriate model is to examine the relationship between each independent variables (topological indices and indicator parameters used in the present study) and the dependent variable lipophilicity (logP) of the compounds used. The correlation matrix (Table 3) is very useful for determining which independent variables are likely to help explain variance in the dependent variable. We look for correlations close to ± 1.0 since that indicates changes in the independent variables are linearly related to changes in the dependent variable.

We can also use correlation matrix to determine the extent to which independent variables are correlated with one another. This can be useful in determining if

Table 1. Set of 140 compounds used in the present investigation, their logP and other molecular descriptors (used for various categories of compounds)

S.N.	Compound	logP	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
1.	Acetylene	0.37	0	0	0	0	0	0	0	0	0
2.	Methylacetylene	0.94	0	0	0	0	0	0	0	0	0
3.	EtOH	−0.31	0	0	0	0	0	0	1	0	0
4.	*i*-Pr–OH	0.05	0	0	0	0	0	0	0	0	0
5.	HCC–CH_2OH	−0.38	0	0	0	0	0	0	1	0	0
6.	*t*-BuOH	0.35	0	0	0	0	0	0	0	0	0
7.	Me–O–Me	0.10	0	0	0	0	0	0	0	0	0
8.	2-Me–oxirane	0.03	0	0	0	0	0	0	0	0	0
9.	2,2-Me_2–1,3-dioxolane	0.46	0	0	0	0	0	0	0	0	0
10.	Et-O–CH_2CH_2OH	−0.35	0	0	0	0	0	0	1	0	0
11.	HS-CH_2–COOH	0.09	1	0	0	0	0	0	0	0	0
12.	AcNH–CH(SH)$CONH_2$	−0.29	0	0	0	0	0	1	0	0	0
13.	5,6-H_2–2Me–1,4-oxathiin–3-COOH	0.04	1	0	0	0	0	0	0	0	0
14.	5,6-H_2–2-Me–1,4-oxathiin–3-$CONH_2$	−0.22	0	0	0	0	0	1	0	0	0
15.	H_2C=O	0.35	0	0	0	0	0	0	0	0	0
16.	Et–CHO	0.59	0	0	0	0	0	0	0	0	0
17.	H_2C=CH–CHO	−0.01	0	0	0	0	0	0	0	0	0
18.	Me–CO–Et	0.29	0	0	0	0	0	0	0	0	0
19.	Ac–CH_2–CH_2–CCH	0.58	0	0	0	0	0	0	0	0	0
20.	cy-Hexanone	0.81	0	0	0	0	0	0	0	0	0
21.	CF_3–CO–Me	0.20	0	0	0	0	0	0	0	0	0
22.	Me–COOH	−0.17	1	0	0	0	0	0	0	0	0
23.	Et–COOH	0.33	1	0	0	0	0	0	0	0	0
24.	Me–CH=CH–COOH	0.72	1	0	0	0	0	0	0	0	0
25.	HOOC–CH_2–CH_2–COOH	−0.59	1	0	0	0	0	0	0	0	0
26.	HOOC–$(CH_2)_3$–COOH	−0.29	1	0	0	0	0	0	0	0	0
27.	Cl–CH_2–COOH	0.22	1	0	0	0	0	0	0	0	0
28.	Me–COOEt	0.73	0	0	0	0	0	0	0	0	0
29.	MeOOC–$(CH_2)_2$COOMe	0.19	0	0	0	0	0	0	0	0	0
30.	EtOOC–COOEt	0.56	0	0	0	0	0	0	0	0	0
31.	CH_2=CH–CH_2–NH_2	0.07	0	0	0	0	0	0	0	0	0
32.	Me_3N	0.16	0	0	0	0	0	0	0	0	0
33.	CF_3CH_2NHEt	0.74	0	0	0	0	0	0	0	0	0
34.	Pyrrolidine	0.46	0	0	0	0	0	0	0	0	0
35.	Et–NH–*i*–Pr	0.93	0	0	0	0	0	0	0	0	0
36.	CH_2=CH–$CONH_2$	−0.67	0	0	0	0	0	1	0	0	0
37.	*n*–Pr–$CONH_2$	−0.21	0	0	0	0	0	1	0	0	0
38.	F–CH_2–$CONH_2$	−1.05	0	0	0	0	0	1	0	0	0
39.	Cl–CH_2–$CONH_2$	−0.53	0	0	0	0	0	1	0	0	0
40.	Br–CH_2–$CONH_2$	−0.52	0	0	0	0	0	1	0	0	0
41.	I–CH_2–$CONH_2$	−0.19	0	0	0	0	0	1	0	0	0
42.	MeOOC–NHMe	−0.06	0	0	0	0	0	0	0	0	0
43.	AcNHCH(CH_2OH)–$CONH_2$	−1.87	0	0	0	0	0	1	1	0	0
44.	Me–Cl	0.91	0	0	0	0	0	0	0	0	0
45.	F–CH_2–CH_2–OH	−0.76	0	0	0	0	0	0	1	0	0
46.	Cl–CH_2–CH_2–OH	−0.06	0	0	0	0	0	0	1	0	0
47.	Br–CH_2–CH_2–OH	0.18	0	0	0	0	0	0	1	0	0
48.	F_2C=CH_2	1.24	0	0	0	0	0	0	0	0	0
49.	Cl–CH_2–CN	0.45	0	0	0	0	0	0	0	0	0
50.	ClCH_2CH_2CN	0.18	0	0	0	0	0	0	0	0	0
51.	Et–NO_2	0.18	0	0	0	0	0	0	0	0	0
52.	*t*-BuNO_2	1.17	0	0	0	0	0	0	0	0	0
53.	Et–CN	0.16	0	0	0	0	0	0	0	0	0
54.	*i*-Pr–CN	0.46	0	0	0	0	0	0	0	0	0
55.	CH_2=CH–CN	0.25	0	0	0	0	0	0	0	0	0
56.	CH_2=CH–CH_2–CN	0.40	0	0	0	0	0	0	0	0	0
57.	Ph–COOH	1.87	1	1	0	0	0	0	0	0	0
58.	Ph–Me	2.73	0	1	0	0	0	0	0	0	0
59.	Ph–O–Me	2.11	0	1	0	0	0	0	0	0	0
60.	Ph–COOMe	2.12	0	1	0	0	0	0	0	0	0
61.	Ph–CH=CH_2	2.95	0	1	0	0	0	0	0	0	0
62.	Ph–CCH	2.53	0	1	0	0	0	0	0	0	0
63.	Ph–CH_2–OH	1.10	0	1	0	0	0	0	1	0	0
64.	Ph–NO_2	1.85	0	1	0	0	0	0	0	0	0
65.	Ph–SH	2.52	0	1	0	0	0	0	0	0	0
66.	Ph–F	2.27	0	1	0	0	0	0	0	0	0
67.	Ph–Cl	2.89	0	1	0	0	0	0	0	0	0
68.	2-F–phenol	1.71	0	1	1	0	0	0	0	0	0
69.	3-F–phenol	1.93	0	1	1	0	0	0	0	0	0
70.	4-F–phenol	1.77	0	1	1	0	0	0	0	0	0
71.	2-Cl–phenol	2.15	0	1	1	0	0	0	0	0	0
72.	3-Cl–phenol	2.50	0	1	1	0	0	0	0	0	0

(continued on next page)

Table 1 (*continued*)

S.N.	Compound	logP	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
73.	4-Cl–phenol	2.39	0	1	1	0	0	0	0	0	0
74.	2-Br–phenol	2.35	0	1	1	0	0	0	0	0	0
75.	3-Br–phenol	2.63	0	1	1	0	0	0	0	0	0
76.	4-Br–phenol	2.59	0	1	1	0	0	0	0	0	0
77.	2-I–phenol	2.65	0	1	1	0	0	0	0	0	0
78.	2-F–aniline	1.26	0	1	0	1	0	0	0	0	0
79.	3-F–aniline	1.30	0	1	0	1	0	0	0	0	0
80.	4-F–aniline	1.15	0	1	0	1	0	0	0	0	0
81.	2-Cl–anilie	1.90	0	1	0	1	0	0	0	0	0
82.	3-Cl–aniline	1.88	0	1	0	1	0	0	0	0	0
83.	4-Cl–aniline	1.88	0	1	0	1	0	0	0	0	0
84.	2-Br–aniline	2.11	0	1	0	1	0	0	0	0	0
85.	3-Br–aniline	2.10	0	1	0	1	0	0	0	0	0
86.	4-Br–aniline	2.26	0	1	0	1	0	0	0	0	0
87.	2-I–aniline	2.32	0	1	0	1	0	0	0	0	0
88.	3-Cl–4-F-aniline	2.06	0	1	0	0	0	0	0	0	0
89.	2-NO_2–phenol	1.79	0	1	1	0	0	0	0	0	0
90.	3-NO_2–phenol	2.00	0	1	1	0	0	0	0	0	0
91.	4-NO_2–phenol	1.91	0	1	1	0	0	0	0	0	0
92.	1-F–2-NO_2–benzene	1.69	0	1	0	0	0	0	0	0	0
93.	1-F–3-NO_2–benzene	1.90	0	1	0	0	0	0	0	0	0
94.	1-F–4-NO_2–benzene	1.80	0	1	0	0	0	0	0	0	0
95.	2-NO_2–aniline	1.85	0	1	0	1	0	0	0	0	0
96.	3-NO_2–aniline	1.37	0	1	0	1	0	0	0	0	0
97.	4-NO_2–aniline	1.39	0	1	0	1	0	0	0	0	0
98.	2-Ac–phenol	1.92	0	1	1	0	0	0	0	0	0
99.	2-Me–benzoicacid	2.46	1	1	0	0	0	0	0	0	0
100.	4-Me–benzoicacid	2.27	1	1	0	0	0	0	0	0	0
101.	Furon	1.30	0	0	0	0	0	0	0	0	1
102.	Pyrrol	0.75	0	0	0	0	0	0	0	0	0
103.	Thiophene	1.81	0	0	0	0	0	0	0	1	0
104.	2-CHO–pyrrole	0.64	0	0	0	0	0	0	0	0	0
105.	2-Ac–pyrrole	0.93	0	0	0	0	0	0	0	0	0
106.	2-COOH–thiophene	1.57	1	0	0	0	0	0	0	1	0
107.	3-COOH–furon	1.03	1	0	0	0	0	0	0	0	1
108.	3-COOH–thiophene	1.50	1	0	0	0	0	0	0	1	0
109.	2-COOMe–furon	1.00	0	0	0	0	0	0	0	0	1
110.	3-COOMe–thiophene	1.76	0	0	0	0	0	0	0	1	0
111.	2-$CONH_2$–furon	−0.11	0	0	0	0	0	1	0	0	1
112.	3-$CONH_2$–furon	0.09	0	0	0	0	0	1	0	0	1
113.	2-CH_2OH–Furon	0.28	0	0	0	0	0	0	1	0	1
114.	2-CH_2OH–thiophene	0.87	0	0	0	0	0	0	1	1	0
115.	3-CH_2OH–furon	0.30	0	0	0	0	0	0	1	0	1
116.	2-Me-O–furon	1.44	0	0	0	0	0	0	0	0	1
117.	2-Me–thiophene	2.33	0	0	0	0	0	0	0	1	0
118.	3-Me–thiophene	2.34	0	0	0	0	0	0	0	1	0
119.	2-Cl-thiophene	2.54	0	0	0	0	0	0	0	1	0
120.	2-Br–thiophene	2.75	0	0	0	0	0	0	0	1	0
121.	2-CN–furon	0.96	0	0	0	0	0	0	0	0	1
122.	2-NO_2–thiophene	1.55	0	0	0	0	0	0	0	1	0
123.	3-NO_2–thiophene	1.55	0	0	0	0	0	0	0	1	0
124.	2-CH=CH-NO_2–thiophene	1.96	0	0	0	0	0	0	0	1	0
125.	Pyridine	0.65	0	0	0	0	1	0	0	0	0
126.	2-Me–Pyridine	1.11	0	0	0	0	1	0	0	0	0
127.	3-Me–Pyridine	1.20	0	0	0	0	1	0	0	0	0
128.	3-OH–Pyridin0e	0.48	0	0	0	0	1	0	0	0	0
129.	2-Me-O–Pyridine	1.36	0	0	0	0	1	0	0	0	0
130.	4-Me-O–Pyridine	1.00	0	0	0	0	1	0	0	0	0
131.	4-CH_2OH–Pyridine	0.06	0	0	0	0	1	0	1	0	0
132.	4-CHO-Pyridine	0.43	0	0	0	0	1	0	0	0	0
133.	2-NH_2–Pyridine	0.49	0	0	0	0	1	0	0	0	0
134.	4-NH_2–Pyridine	0.32	0	0	0	0	1	0	0	0	0
135.	2-$CONH_2$-Pyridine	0.15	0	0	0	0	1	1	0	0	0
136.	3-$CONH_2$–Pyridine	−0.34	0	0	0	0	1	1	0	0	0
137.	4-$CONH_2$-Pyridine	−0.28	0	0	0	0	1	1	0	0	0
138.	2-NH_2–3-$CONH_2$–pyridine	0.88	0	0	0	0	1	1	0	0	0
139.	3-CN–pyridine	0.23	0	0	0	0	1	0	0	0	0
140.	4-NO_2–pyridine	0.33	0	0	0	0	1	0	0	0	0

logP, octanol/water partition coefficient allocating for lipophilicity of the compounds used; M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8 and M_9 are the molecular descriptors used for the presence of (1) or absence (0) of –COOH–C_6H_5, monosubstituted phenols, monosubstituted anilines, pyridine nucleous, –$CONH_2$, –CH_2OH, thiophene, and furan, respectively.

Table 2. Topological indices calculated for the set of 140 compounds used in the present study (ref. Table 1)

Compd	W	B	χ	J	Sz	log(RB)	W*
1	1	1.0000	1.0000	1.0000	1	0.0000	1
2	4	1.4142	1.4142	1.6329	4	0.6931	5
3	4	1.4142	1.4142	1.6329	4	0.9631	5
4	9	1.7320	1.7320	2.3237	9	2.0794	12
5	10	1.9142	1.9142	1.9747	10	2.4849	15
6	16	2.0000	2.0000	3.0237	16	4.1589	22
7	4	1.4142	1.4142	1.6329	4	0.6931	5
8	8	1.8938	1.8938	2.1711	8	1.3863	10
9	39	3.2071	3.2071	2.3997	48	11.3259	61
10	35	2.9142	2.9142	2.3390	35	10.4505	70
11	18	2.2700	2.2700	2.5395	18	4.9698	28
12	136	4.5366	4.5366	3.3014	136	42.6536	330
13	114	4.7152	4.7152	2.3960	180	36.5035	231
14	114	4.7152	4.7152	2.3960	180	36.5035	231
15	1	1.0000	1.0000	1.0000	1	0.0000	1
16	10	1.9142	1.9142	1.9747	10	2.4849	15
17	10	1.9142	1.9142	1.9747	10	2.4849	15
18	18	2.2700	2.2700	2.5395	18	4.9698	28
19	79	3.7701	3.7701	2.7158	79	24.3021	185
20	114	4.6051	4.6051	2.3892	177	36.3212	233
21	42	2.9433	2.9433	3.5412	42	12.8300	69
22	9	1.7320	1.7320	2.3237	9	2.0794	12
23	18	2.2700	2.2700	2.5395	18	4.9698	28
24	32	2.7000	2.7000	2.6272	32	9.5342	58
25	74	3.6259	3.6259	2.9278	74	23.0212	161
26	108	4.1259	4.1259	2.9146	108	33.3663	265
27	18	2.2700	2.2700	2.5395	18	4.9698	28
28	32	2.7700	2.7700	2.6272	32	9.5342	58
29	141	4.7019	4.7019	3.1681	141	43.9063	355
30	135	4.7187	4.7187	3.3759	135	42.5846	324
31	10	1.9142	1.9142	1.9747	10	2.4849	15
32	9	1.7320	1.7320	2.3237	9	2.0794	12
33	71	3.5606	3.5606	3.1117	71	22.1049	149
34	15	2.5000	2.5000	2.0833	20	3.4657	20
35	32	2.7700	2.7700	2.6272	32	9.5342	58
36	18	2.2700	2.2700	2.5395	18	4.9698	28
37	32	2.7700	2.7700	2.6272	32	9.5342	58
38	18	2.2700	2.2700	2.5395	18	4.9698	28
39	18	2.2700	2.2700	2.5395	18	4.9698	28
40	18	2.2700	2.2700	2.5395	18	4.9698	28
41	18	2.2700	2.2700	2.5395	18	4.9698	28
42	31	2.8080	2.8080	2.7541	31	9.2465	54
43	174	5.0746	5.0746	3.5226	174	54.8958	433
44	1	1.0000	1.0000	1.0000	1	0.0000	1
45	10	1.9142	1.9142	1.9747	10	2.4849	15
46	10	1.9142	1.9142	1.9747	10	2.4849	15
47	10	1.9142	1.9142	1.9747	10	2.4849	15
48	9	1.7320	1.7320	2.3237	9	2.0794	12
49	10	1.9142	1.9142	1.9747	10	2.4849	15
50	20	2.4142	2.4142	2.1906	20	5.663	35
51	18	2.2700	2.2700	2.5395	18	4.9698	28
52	52	3.2700	3.2700	2.6782	52	15.9311	108
53	10	1.9142	1.9142	1.9747	10	2.4849	15
54	18	2.2700	2.2700	2.5395	18	4.9698	28
55	4	1.4142	1.4142	1.6329	4	0.6931	5
56	20	2.4142	2.4142	2.1906	20	5.6630	35
57	88	4.3045	4.3045	2.2283	142	27.6625	176
58	42	3.3938	3.3938	2.1229	78	12.4245	71
59	64	3.9318	3.9318	2.1250	109	19.6969	122
60	121	4.8425	4.8425	2.2395	184	38.3361	263
61	64	3.9318	3.9318	2.1250	109	19.6969	122
62	64	3.9318	3.9318	2.1250	109	19.6969	122
63	64	3.9318	3.9318	2.1250	109	19.6969	122
64	88	4.3045	4.3045	2.2283	142	27.6625	176
65	42	3.3938	3.3938	2.1229	78	12.4245	71
66	42	3.3938	3.3938	2.1229	78	12.4245	71
67	42	3.3938	3.3938	2.1229	78	12.4245	71
68	60	3.8045	3.8045	2.2794	106	18.493	106
69	61	3.7876	3.7876	2.2306	108	18.7806	110
70	62	3.7876	3.7876	2.1923	110	19.0038	115
71	60	3.8045	3.8045	2.2794	106	18.493	106
72	61	3.7876	3.7876	2.2306	108	18.7806	115

(continued on next page)

Table 2 (*continued*)

Compd	W	B	χ	J	Sz	log(RB)	W*
73	62	3.7876	3.7876	2.1923	110	19.0038	115
74	60	3.8045	3.8045	2.2794	106	18.493	106
75	61	3.7876	3.7876	2.2306	108	18.7806	110
76	62	3.7876	3.7876	2.1923	110	19.0038	115
77	60	3.8045	3.8045	2.2794	106	18.493	106
78	60	3.8048	3.8048	2.2794	106	18.493	106
79	61	3.7876	3.7876	2.2306	108	18.7806	110
80	62	3.7876	3.7876	2.1923	110	19.0038	115
81	60	3.8045	3.8045	2.2794	106	18.493	106
82	61	3.7876	3.7876	2.2306	108	18.7806	110
83	62	3.7876	3.7876	2.1923	110	19.0038	115
84	60	3.8045	3.8045	2.2794	106	18.493	106
85	61	3.7876	3.7876	2.2306	108	18.7806	110
86	62	3.7876	3.7876	2.1923	110	19.0039	115
87	60	3.8045	3.8045	2.2794	106	18.493	106
88	84	4.1983	4.1983	2.3462	144	26.4585	160
89	114	4.7152	4.7152	2.3960	180	36.5035	231
90	117	4.6983	4.6983	2.3198	186	37.2374	245
91	120	4.6983	4.6983	2.2599	192	37.8252	262
92	114	4.7152	4.7152	2.3960	180	36.5035	231
93	117	4.6983	4.6983	2.3198	186	37.2374	245
94	120	4.6983	4.6983	2.2599	192	37.8252	262
95	114	4.7152	4.7152	2.3960	180	36.5035	231
96	117	4.6983	4.6983	2.3198	186	37.2374	245
97	120	4.6983	4.6983	2.2599	192	37.8252	262
98	114	4.7152	4.7152	2.3960	180	36.5053	231
99	114	4.7152	4.7152	2.3960	180	36.5035	231
100	120	4.6983	4.6983	2.2598	192	37.8252	262
101	15	2.5000	2.5000	2.0833	20	3.4657	20
102	15	2.5000	2.5000	2.0833	20	3.4657	20
103	15	2.5000	2.5000	2.0833	20	3.4657	20
104	43	3.4318	3.4318	2.1399	52	12.7122	75
105	93	4.2877	4.2877	2.1128	106	28.9718	198
106	62	3.8045	3.8045	2.2420	73	19.0683	114
107	62	3.8045	3.8045	2.2420	73	19.0683	114
108	62	3.8045	3.8045	2.2420	73	19.0683	114
109	89	4.3425	4.3425	2.2354	102	27.9502	180
110	89	4.3425	4.3425	2.2354	102	27.9502	180
111	62	3.8045	3.8045	2.2420	73	19.0683	114
112	62	3.8045	3.8045	2.2420	73	19.0683	114
113	43	3.4318	3.4318	2.1399	52	12.7122	75
114	43	3.4318	3.4318	2.1399	52	12.7122	75
115	43	3.4318	3.4318	2.1399	52	12.7122	75
116	43	3.4318	3.4318	2.1399	52	12.7122	75
117	26	2.8938	2.8938	2.1841	33	7.0493	39
118	26	2.8938	2.8938	2.1841	33	7.0493	39
119	26	2.8938	2.8938	2.1841	33	7.0493	39
120	26	2.8938	2.8938	2.1841	33	7.0493	39
121	43	3.4318	3.4318	2.1399	52	12.7122	75
122	62	3.8045	3.8045	2.2420	73	19.0683	114
123	62	3.8045	3.8045	2.2420	73	19.0683	114
124	133	4.7876	4.7876	2.0127	148	41.1495	322
125	27	3.0000	3.0000	2.0000	54	7.4547	42
126	42	3.3938	3.3938	2.1229	78	12.4245	71
127	42	3.3938	3.3938	2.1229	78	12.4245	71
128	42	3.3938	3.3938	2.1229	78	12.425	71
129	64	3.9318	3.9318	2.1250	109	19.6969	122
130	64	3.9318	3.9318	2.1250	109	19.6969	122
131	64	3.9318	3.9318	2.1250	109	19.6969	122
132	64	3.9318	3.9318	2.1250	109	19.6969	122
133	42	3.3938	3.3938	2.1229	78	12.4245	71
134	42	3.3938	3.3938	2.1229	78	12.4245	71
135	88	4.3045	4.3045	2.2283	142	27.6625	176
136	88	4.3045	4.3045	2.2283	142	27.6625	176
137	88	4.3045	4.3045	2.2283	142	27.6625	176
138	114	4.7152	4.7152	2.3960	180	36.5035	231
139	64	3.9318	3.9318	2.1250	109	19.6969	122
140	88	4.3045	4.3045	2.2283	142	27.6625	176

W, Wiener index; B, branching index; χ, connectivity index; J, Balaban index; Sz, Szeged index, W*, hyper Wiener index; logRB, type of branching index.

Table 3. Correlation matrix for the correlation of various molecular descriptors

	logP	W	B	χ	J	Sz	log(RB)	W*	
logP	1.0000								
W	0.2421	1.0000							
B	0.4011	0.9287	1.0000						
X	0.4011	0.9287	1.0000	1.0000					
J	−0.1983	0.4050	0.3683	0.3683	1.0000				
Sz	0.4030	0.9377	0.9398	0.9398	0.2339	1.0000			
log(RB)	0.2461	0.9952	0.9327	0.9326	0.4014	0.94664	1.0000		
W*	0.1799	0.9873	0.8872	0.8872	0.4284	0.9030	0.9898	1.0000	
M_1	−0.0627	0.0787	0.0458	0.0458	0.1710	0.0274	0.0778	0.0790	
M_2	0.7264	0.3950	0.4883	0.4883	−0.0529	0.5932	0.4110	0.3469	
M_3	0.3957	0.1842	0.2415	0.2415	−0.0093	0.3025	0.2112	0.1799	
M_4	0.2479	0.1545	0.2052	0.2052	−0.0111	0.2484	0.1533	0.1154	
M_5	−0.2710	−0.1207	−0.1456	−0.1456	0.2140	−0.1804	−0.1269	−0.0954	
M_6	−0.4207	0.2006	0.1616	0.1616	0.2098	0.1477	0.2001	0.2022	
M_7	−0.3234	−0.0980	−0.1236	−0.1236	−0.0834	−0.1453	−0.1009	−0.0761	
M_8	0.2758	−0.0182	0.0419	0.0419	−0.0755	−0.0937	−0.0281	−0.0274	
M_9	−0.0730	−0.0246	0.0527	0.0527	−0.0668	−0.0854	−0.0306	−0.0443	
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
M_1	1.0000								
M_2	−0.0718	1.0000							
M_3	−0.1111	0.4924	1.0000						
M_4	−0.1067	0.4726	−0.1067	1.0000					
M_5	−0.1197	−0.2432	−0.1197	−0.1149	1.0000				
M_6	−0.1155	−0.2345	−0.1155	−0.1108	0.0933	1.0000			
M_7	−0.1021	−0.1523	−0.1021	−0.0980	−0.1100	−0.0236	1.0000		
M_8	0.0680	−0.2073	−l0.1021	−0.0980	−0.1100	−0.1061	−0.0026	1.0000	
M_9	0.0097	−0.1775	−0.0874	−0.0839	−0.0942	0.0975	0.1278	−0.0803	1.0000

certain independent variables are redundant and not needed in the model. The correlation matrix (Table 3) has 1.0000 running down the main diagonal, indicating the variables are perfectly correlated with themselves.

A perusal of Table 3 shows that none of the topological index correlate significantly with the lipophilicity (logP) meaning thereby that no mono-parametric models are possible for modelling the lipophilicity of the compounds used. Only the molecular descriptor M_2 correlates well with the lipophilicity. As is shown in the experimental section below, M_2 stands for the presence of $-C_6H_5$ moiety. Thus, correlation of M_2 with logP indicates that presence of $-C_6H_5$ moiety is a pre-requisite for the exhibition of lipophilicity.

The correlation matrix (Table 3) also show high collinearity between: (i) W, B; (ii) W, χ (iii) W, Sz; (iv) W,

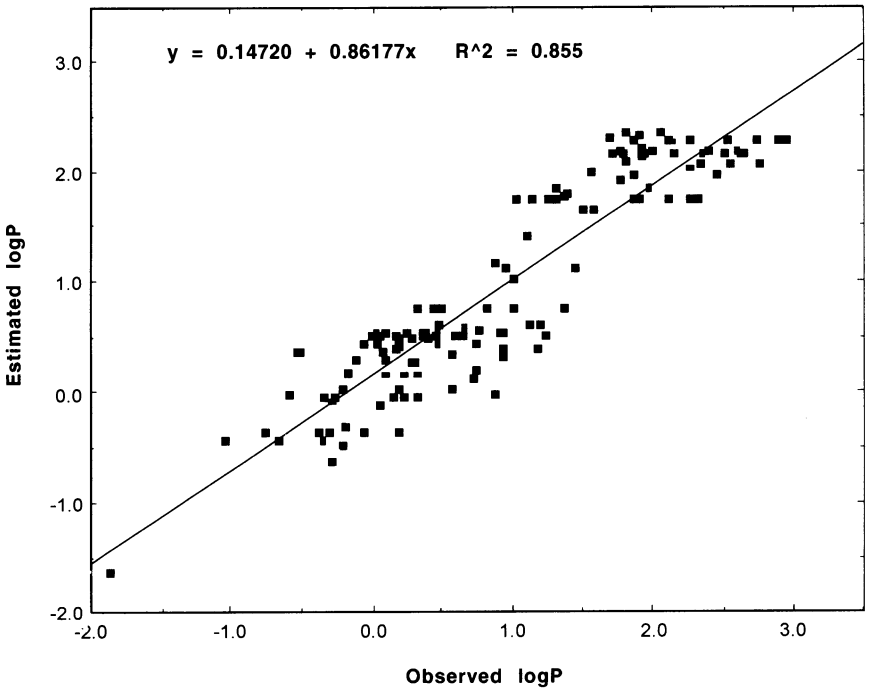

Figure 1. Comparison of observed and estimated lipophilicity (logP) using model 8.

Table 4. Regression parameters and quality of the proposed models

Model no.	Parameters used	Eq. no.	Coefficients (A_i) $i=9$, 10	Constant B	Se	R^2	R	F ratio	$Q=R/Se$
1	W		0.0141 (±0.0033)	0.4829	0.4079	0.8453	0.9194	78.941	2.254
	logRB		0.0460 (±0.0290)						
	M_1		−0.3277 (±0.1194)						
	M_2		1.7921 (±0.1040)						
	M_4		−0.5097 (±0.1362)						
	M_6		−0.7223 (±0.1267)						
	M_7		−0.8867 (±0.1267)						
	M_8		1.5060 (±0.1299)						
	M_9		0.5972 (±0.1451)						
2	Sz		9.3469×10^{-4} ($\pm8.1719\times10^{-4}$)	1.006	0.403	0.849	0.9214	81.206	2.2863
	J		−0.2431 (±0.1022)						
	M_1		−0.3012 (±0.1192)						
	M_2		1.6446 (±0.1161)						
	M_4		−0.4678 (±0.1340)						
	M_6		−0.7509 (±0.1271)						
	M_7		−0.8963 (±0.1254)						
	M_8		1.4684 (±0.1285)						
	M_9		0.5648 (±0.1439)						
3	W		−0.0089 (±0.0031)	0.5011	0.3993	0.8517	0.9229	82.988	2.3113
	Sz		0.0068 (±0.0024)						
	M_1		−0.3127 (±0.1171)						
	M_2		1.5224 (±0.1286)						
	M_4		−0.4734 (±0.1327)						
	M_6		−0.7936 (±0.1252)						
	M_7		−0.8536 (±0.1246)						
	M_8		1.5443 (±0.1274)						
	M_9		0.6393 (±0.1424)						
4	W		−0.0092 (±0.0031)	0.5049	0.3988	0.8533	0.9237	75.016	2.3162
	Sz		0.0070 (±0.0024)						
	M_1		−0.3317 (±0.1181)						
	M_2		1.5908 (±0.1414)						
	M_3		−0.1689 (±0.1461)						
	M_4		−0.5516 (±0.1988)						
	M_6		−0.7979 (±0.1251)						
	M_7		−0.8626 (±0.1247)						
	M_8		1.5458 (±0.1272)						
	M_9		0.6419 (±0.1423)						
5	W		−0.0080 (±0.0026)	−0.1283	0.3967	0.8537	0.924	84.282	2.3381
	B		0.3318 (±0.1013)						
	M_1		−0.3378 (±0.1113)						
	M_2		1.5780 (±0.1045)						
	M_4		−0.4851 (±0.1255)						
	M_6		0.8553 (±0.1100)						
	M_7		−0.8965 (±0.1178)						
	M_8		1.3514 (±0.1274)						
	M_9		0.4361 (±0.1429)						
6	W		0.0197 (±0.0061)	0.3913	0.3954	0.8547	0.9245	84.963	2.3381
	W*		−0.0086 (±0.0026)						
	M_1		−0.3476 (±0.1158)						
	M_2		1.6576 (±0.1029)						
	M_4		−0.5403 (±0.1132)						
	M_6		−0.7830 (±0.1232)						
	M_7		−0.8419 (±0.1235)						
	M_8		1.4408 (±0.1278)						
	M_9		0.4984 (±0.1442)						
7	W		−0.0080 (±0.0025)	−0.1315	0.3966	0.8549	0.9246	76.023	2.3313
	B		0.3382 (±0.1053)						
	M_1		−0.3349 (±0.1174)						
	M_2		1.6730 (±0.1264)						
	M_3		−0.1521 (±0.1450)						
	M_4		−l0.5570 (±0.1429)						
	M_6		−0.7829 (±0.1236)						
	M_7		−0.8636 (±0.1239)						
	M_8		1.3760 (±0.1335)						
	M_9		0.4399 (±0.1502)						
8	Sz	1	0.0093 (±0.0025)	0.4767	0.3913	0.8577	0.9261	87.046	2.3667
	logRB		−0.0382 (±0.0102)						
	M_1		−0.2983 (±0.1148)						
	M_2		1.4769 (±0.1255)						
	M_4		−0.4890 (±0.1300)						
	M_6		0.7925 (±0.1227)						

(continued on next page)

Table 4 (*continued*)

Model no.	Parameters used	Eq. no.	Coefficients (A_i) $i=9$, 10	Constant B	Se	R^2	R	F ratio	$Q=R/Se$
	M_7		−0.8432 (±0.1221)						
	M_8		1.5490 (±0.1246)						
	M_9		0.6458 (±0.1395)						
9	Sz		0.0067 (±0.0019)	0.4548	0.3921	0.8582	0.9264	78.053	2.3627
	W*		−0.0038 (±0.0010)						
	M_1		−0.3340 (±0.1159)						
	M_2		1.5526 (±0.1391)						
	M_3		−0.1436 (±0.1433)						
	M_4		−0.5648 (±0.1463)						
	M_6		−0.8092 (±0.1231)						
	M_7		0.8433 (±0.1229)						
	M_8		1.5132 (±0.1249)						
	M_9		0.5936 (±0.1393)						
10	Sz	2	0.0089 (±0.0025)	0.5107	0.3908	0.8591	0.9269	78.649	2.3718
	logRB		−0.0368 (±0.0103)						
	M_1		−0.3201 (±0.1163)						
	M_2		1.4577 (±0.1265)						
	M_4		−0.4917 (±0.1298)						
	M_5		−0.1293 (±0.1136)						
	M_6		−0.7957 (±0.1225)						
	M_7		−0.8684 (±0.1240)						
	M_8		1.5206 (±0.1269)						
	M_9		0.6196 (±0.1412)						
11	Sz	3	0.0089 (±0.0025)	0.5145	0.3908	0.8602	0.9275	71.6	2.3733
	logRB		−0.0368 (±0.0103)						
	M_1		−0.3378 (±0.1176)						
	M_2		1.5198 (±0.1407)						
	M_3		−0.1440 (±0.1429)						
	M_4		−0.5587 (±0.1458)						
	M_5		−0.1329 (±0.1137)						
	M_6		−0.7956 (±0.1226)						
	M_7		−0.8775 (±0.1243)						
	M_8		1.5201 (±0.1269)						
	M_9		0.6201 (±0.1412)						

A_i, B, regression parameters; R, correlation coefficient; F, Fischer's ratio; Se, standard error of estimation; Q, quality factor.

logRB; (v) W, W*; (vi) B, Sz; (vii) B, W*; (viii) χ, Sz; (ix) χ, W*; (x) Sz, logRB, slightly lower collinearity exists between: (i) B, logRB; (ii) χ, logRB; (iii) Sz, logRB; (iv) Sz, Ip_2; (v) logRB, W*. This shows that the former pairs of topological indices may cause collinearity defects if they are present in the model as correlating parameters. It is worth recording that the topological indices J, W* and the other molecular descriptors do not correlate with remaining topological indices used. As stated earlier only M_2 correlates well with logP. This means that J, W*, M_2 are the most appropriate molecular descriptors to be used in obtaining statistically significant multi-parametric models.

As stated earlier our main objective of the present investigation is to provide topological evidences for modelling lipophilicity (logP) in that we are mainly concerned with the relative potential of Szeged index for this purpose. We have used maximum R^2 improvement method[35] for this purpose and observed that fruitful models are obtained only when 9, 10 and 11 parameters are involved in the regression model. Certainly more than two topological indices gave better results as shown in Tables 4 and 5, The models obtained are the combinations of Sz or W with some other topological index along with a set of other molecular descriptors. Statistically significant models were also obtained when both W and Sz are present simultaneously.

Table 5. Summary table of the statistics used

Regression expression	Parameters used	Se	R^2	R	F	Q
1	W, logRB, M_1, M_2, M_4, M_6, M_7, M_8, M_9	0.4079	0.8453	0.9194	78.941	2.2540
2	Sz, J, M_1, M_2, M_4, M_6, M_7, M_8, M_9	0.4030	0.8490	0.9214	81.206	2.2863
3	W, Sz, M_1, M_2, M_4, M_6, M_7, M_8, M_9	0.3993	0.8517	0.9229	82.988	2.3113
4	W, Sz, M_1, M_2, M_3, M_4, M_6, M_7, M_8, M_9	0.3988	0.8533	0.9237	75.016	2.3162
5	W, B, M_1, M_2, M_4, M_6, M_7, M_8, M_9	0.3967	0.8537	0.9240	84.282	2.3381
6	W, W*, M_1, M_2, M_4, M_6, M_7, M_8, M_9,	0.3954	0.8547	0.9245	84.963	2.3381
7	W, B, M_1, M_2, M_3, M_4, M_6, M_7, M_8, M_9	0.3966	0.8549	0.9246	76.023	2.3313
8	Sz, logRB, M_1, M_2, M_4, M_6, M_7, M_8, M_9	0.3913	0.8577	0.9261	87.046	2.3667
9	Sz, W*, M_1, M_2, M_3, M_4, M_6, M_7, M_8, M_9	0.3921	0.8582	0.9264	78.053	2.3627
10	Sz, logRB, M_1, M_2, M_4, M_5, M_6, M_7, M_8, M_9	0.3908	0.8591	0.9269	78.649	2.3718
11	Sz, logRB, M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9	0.3908	0.8602	0.9275	71.600	2.3733

Table 6. Cross-validation parameters and root mean square error of residues (*RMSR*) for the statistically most significant models

Model	No of descriptors involved	R^2_A	VIF	PRESS	SSY	PRESS/SSY	R^2_{CV}	S_{PRESS}	S_{PSE}	RMSR
1	9	0.8346	6.4641	21.6299	118.2102	0.1830	0.8170	0.4079	0.3931	—
2	9	0.8385	6.6225	21.1177	118.7223	0.1779	0.8221	0.4030	0.3884	—
3	9	0.8415	6.7431	20.7314	119.1086	0.1740	0.8260	0.3993	0.3848	—
4	10	0.8419	6.8166	20.5789	119.3211	0.1722	0.8278	0.3988	0.3828	—
5	9	0.8436	6.8353	20.4596	119.3805	0.1714	0.8286	0.3967	0.3823	—
6	9	0.8446	6.8823	20.3196	119.5205	0.1784	0.8216	0.3954	0.3810	—
7	10	0.8437	6.8918	20.2864	119.5536	0.1697	0.8303	0.3966	0.3801	—
8 (1)	9	0.8478	7.0274	19.9025	119.9376	0.1659	0.8341	0.3913	0.3770	0.3771
9	10	0.8472	7.0522	19.8337	120.0064	0.1653	0.8347	0.3921	0.3764	—
10 (2)	10	0.8482	7.0972	19.7046	120.1354	0.1640	0.8360	0.3908	0.3752	0.3753
11 (3)	11	0.8482	7.1531	19.5495	120.2906	0.1625	0.8375	3908	0.3736	0.3807

R^2_A- adjustable R^2_A, VIF, variance inflation factor; PRESS, predicted residual sum of squares; SSY, sum of the squares of the response value; R^2_{CV}, cross-validation correlation coefficient; S_{PRESS}, uncertainty of prediction; S_{PSE}, predictive squre error, RMSE, root mean square error of the residue. Values in the paranthesis in column 1 shows the regression equations used in the text.

A perusal of Table 5 shows that there exists six statistically significant regression expressions involving nine molecular descriptors. Out of these nine descriptors two are topological indices and the remaining seven are the molecular descriptors related to the type of the compounds used. In each of these regressions Sz or W is the main and commendable topological index. The statistical parameters (Se, R, R^2, F and Q) indicate that out of the six nine-parametric regressions, the regression expression containing Sz, logRB, M_1, M_2, M_4, M_6, M_7, M_8 and M_9 is the most significant for modelling the lipophilicity (logP) of the compounds used. This model is found as under:

$$\begin{aligned}\log P &= 0.0093(\pm 0.0025)Sz - 0.0382 \\ &\times (\pm 0.0102)\log RB - 0.2983(\pm 0.1148)M_1 \\ &+ 1.4769(\pm 0.1255)M_2 - 0.4890 \\ &\times (\pm 0.1300)\ M_4 + 0.7925 \\ &\times (\pm 0.1227)M_6 - 0.8432(\pm 0.1221)M_7 \\ &+ 1.5490(\pm 0.1246)M_8 + 0.6458 \\ &\times (\pm 0.1395)\ M_9 + 0.4767 \\ n &= 140,\ \mathrm{ND} = 9,\ \mathrm{Se} = 0.3913,\ R = 0.9261, \\ F &= 87.046, Q = 2.3667 \end{aligned} \quad (1)$$

The data presented in Table 5 also shows the occurrence of four statistically significant regression expressions containing 10 molecular descriptors. Out of these four expressions only two gave slightly better results than the nine-parametric model discussed above. These two models are found to contain (i) Sz, W*, M_1, M_2, M_3, M_4, M_6, M_7, M_8, M_9 and (ii) Sz, logRB, M_1, M_2, M_4, M_5, M_6, M_7, M_8 and M_9, respectively. Here these ten parametric models are obtained by introducing additional molecular descriptors M_3 and M_5, respectively. These molecular descriptors are responsible for the presence of phenolic and pyridine moieties, respectively. This means that introduction of phenolic and pyridine compounds in the set of 140 compounds slightly improves the statistics so that *R*-value increases from 0.9261 (nine-parametric model) to 0.9264 and 0.9269 (10-parametric models). Out of the two 10-parametric models the model containing Sz, logRB, M_1, M_2, M_4, M_5, M_6, M_7, M_8 and M_9 is found to be the better and is found as below:

$$\begin{aligned}\log P &= 0.0089(\pm 0.0025)Sz - 0.0368(\pm 0.0103)\log RB \\ &- 0.3201(\pm 0.1163)M_1 + 1.4577 \\ &\times (\pm 0.1265)M_2 - 0.4917 \\ &\times (\pm 0.1298)M_4 - 0.1293(\pm 0.1136)M_5 \\ &- 0.7957(\pm 0.1225)M_6 - 0.8684 \\ &\times (\pm 0.1240)M_7 + 1.5206(\pm 0.1269)M_8 \\ &+ 0.6196(\pm 0.1412)M_9\ + 0.5107 \\ n &= 140,\ \mathrm{ND} = 10,\ \mathrm{Se} = 0.3908, \\ R &= 0.9269, F = 78.649, Q = 2.3718 \end{aligned} \quad (2)$$

Finally, an 11-parametric model containing Sz, logRB, M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8 and M_9 gave the best

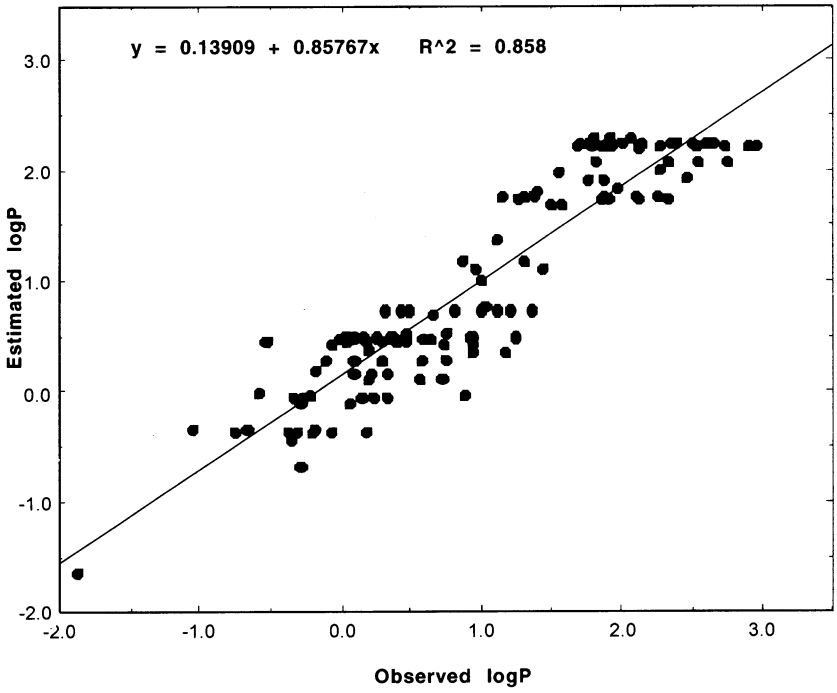

Figure 2. Comparison of observed and estimated lipophilicity (logP) using model 10.

results. This model is found as under

$$\begin{aligned}\log P = {} & 0.0089(\pm 0.0025)Sz - 0.0368 \\ & \times (\pm 0.0103)\ \log RB - 0.3378(\pm 0.1176)M_1 \\ & + 1.5198(\pm 0.1407)M_2 - 0.1440 \\ & \times (\pm 0.1429)M_3 - 0.5587 \\ & \times (\pm 0.1458)M_4 - 0.1329(\pm 0.1137)M_5 - 0.7956 \\ & \times (\pm 0.1226)M_6 - 0.8775(\pm 0.1243)M_7 \\ & + 1.5201(\pm 0.1269)\ M_8 + 0.6201 \\ & \times (\pm 0.1412)\ M_9 + 0.5145 \\ n = 140,\ & \mathrm{ND} = 11, \mathrm{Se} = 0.3908, \\ R = 0.9275,\ & F = 71.600,\ Q = 2.3733 \end{aligned} \quad (3)$$

It is worth mentioning that in all the regression expressions summarized in Table 5 we have used the quality factor (Q) for establishing the quality of the proposed models. This quality factor (Q) is defined[38,39] as the ratio of correlation coefficient (R) to the standard error of estimation (Se), that is, $Q = R/\mathrm{Se}$. Thus, higher the value of R, the lower the Se, the larger will be Q, and better will be the quality of the model. These Q values also suggest that the model expressed by eq 3 is better than the other two models.

The data presented in Tables 4 and 5 show that multiparametric models involving the Sz index gave better results than the models involving the W index. This shows that compared to W, Sz is a better topological index for modelling lipophilicity of the compounds used. Also, that in addition to above we obtained statistically significant 9- and 10-parametric models in that both W and Sz are simultaneously present. As stated earlier W and Sz are highly linearly correlated with each other (Table 3) and their presence should have aroused collinearity defect in these models. While this is not the case, instead in both the cases the respective coefficients of W and Sz were appreciably larger than their standard deviations. Such models are considered statistically significant.[40] Diagnostics to test for the presence of compounds outliers and multi-collinearity between descriptors in the model a variance inflation factor (VIF) is proposed in the literature.[41] This factor less than 10 indicates that the model contained no multicollinearity. The calculated VIF in the present study are presented in Table 6. All the regression expressions (models) have VIF values less than 10 indicating that none of them suffers from collinearity defect.

Randic[42] has stated that one should particularly be aware of a common fit all in regression analysis in describing descriptors that are highly inter-correlated. He further stated that by discarding one of the descriptors which commonly duplicates another we may be discarding a descriptor that nevertheless may carry useful structural information in the parts in which it does not parallel with the another descriptors. Thus, following Randic,[42] we may safely say that in the referred statistically significant models containing both W and Sz simultaneously, the Sz index caries unknown structural information not present in W. Recall that these

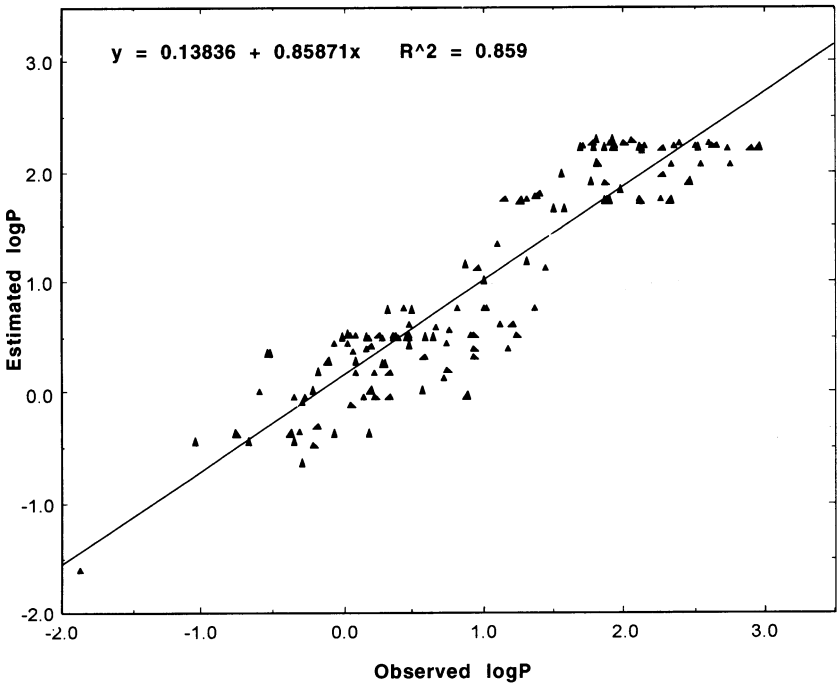

Figure 3. Comparison of observed and estimated lipophilicity (logP) using model 11.

descriptors, namely W and Sz, are highly linearly correlated. A detailed study on this descriptor is underway and will be published elsewhere.

A perusal of Table 6 shows that adjustable R^2, that is, R^2_A value is the highest for the model expressed by eq 3. Adjustable R^2, namely R^2_A is defined as:

$$R^2_A = 1 - \frac{(n-1)(1-R^2)}{(n-\mathrm{ND})}$$

where n = number of compounds used; R^2 = square of correlation coefficient R and ND = number of descriptors involved in the models.

R^2_A is a measure of % explained variance in the dependent variable (in our case logP) that takes into account the relationship between the number of cases (compounds used) and the number of independent variables (topological indices and other molecular descriptors) in the regression model; R^2 will always increase when an independent variable is added; R^2_A on the other hand will decrease if the added variable does not reduce the unexplained variation enough to offset the loss of degrees of freedom. It means that the 10-parametric model expressed by eq 2 is slightly better than one expressed by eq 1.

The aforementioned results and discussion indicates that the lipophilicity (logP) of the set of compounds (140) used is a multidimensional property (ND being 9, 10 and 11) and cannot be expressed by any combination of one dimensional property. Thus, the combinations of two topological indices with 7, 8 or 9 other molecular descriptors (used for various categorries of compounds) indeed describe lipophilicity, and then one is obliged to demonstrate they can also describe other receptor activity with same combination of indices. Thus, application of topological indices and their combinations with other molecular descriptors (used for various categorries of compounds) widens the problem of providing structural evidences of lipophilicity. This is true because a topological approach enables any structure to be described by means of a number index which has physico-chemical meaning. However, it must be said that the range of applications of these indices in direct correlation is limited. They can be applied as the sole parameters only then when the correlated quantity (property or activity) depends upon the following features and elements of molecular structure: volume, length of carbon chain, ring size, plain of chain branching. Hence, results obtained by us indicate that in addition to the type of the compounds used, the parameters therein provide evidence for the modelling of lipophilicity for the set of 140 compounds used.

It is worth while mentioning that good statistics alone do not mean that the corresponding models have better predictive potential also. In view of this, we used a cross-validation method[35] and calculated various cross-validation parameters for judging the predictive potential of all the ten models proposed (Table 6). The cross-validation parameters used being: PRESS (predicted residual sum of squares), SSY (sum of the squares of the response value), R^2_{CV} (cross-validation correlation coefficient), S_{PRESS} (uncertainty of prediction) and S_{PSE} (predictive square error). These parameters as reported in Table 5 are defined by the following expressions:

$$R^2\text{cv} = (\text{SSY} - \text{PRESS})/\text{SSY};$$
$$S_{\text{PRESS}} = [\text{PRESS}/n - \text{ND} - 1]^{0.5}$$
$$S_{\text{PSE}} = [\text{PRESS}/n]^{0.5}$$

where ND is the number of descriptors in the model and n is the total number of compounds used in the analysis. The calculated values of these parameters are given in Table 6.

PRESS is a good estimate of the real prediction error of the model. If PRESS is smaller than the sum of the squares of the response value (SSY), the model predict better than chance can be considered statistically significant. A perusal of the data presented in Table 6 show that this is the case with all the models attempted in the present study. Furthermore, to be a reasonable QSAR/QSPR model, PRESS/SSY should be smaller than 0.4 and value of this ratio smaller than 0.1 indicates an excellent model. In the present case this ratio is found to vary between 0.16 and 0.18 indicating thereby that all the proposed models are statistically excellent, and that model-11 (eq 3) is the most appropriate model.

It is worthy to note that the value of PRESS does not change significantly when the number of descriptors is increased from 8 to 11. It is noteworthy that the lowest PRESS value (19.5495) is obtained for the model represented by (eq 3). This, along with the earlier discussion shows that model expressed by eq 3 has the highest predictive potential.

Another cross-validation parameter used for expressing overall predictive ability of model(s) is R^2_{CV}. The value of R^2_{CV} nearer to 1 indicates highest predictive ability of the model. The R^2_{CV} values as recorded in Table 6 ranges between 0.81 and 0.84, being highest for the model expressed by (eq 3). Once again we obtained an evidence in favour of this model.

Further support in favour of this model could be obtained from the value of S_{PRESS}. Such values are presented in Table 6. Unfortunately, we observed that S_{PRESS} is the same as standard error of estimation, Se. Hence, we can not use S_{PRESS} for accounting predictive potential of the model and thus we have to search for some other better parameter for this purpose. Fortunately, we have predictive squared error (S_{PSE}) for this purpose, which seems to be more directly related to the uncertainty of prediction. The lowest S_{PSE} value (0.3736) for the model expressed by eq 3 (Table 6) indicates it to have the highest predictive potential.

It is worth mentioning that some of the parameters involved in eqs 1 and 2 have positive coefficient and some have negative coefficients. This means that in some cases lipophilicity increases with the magnitude of the parameters with positive coefficients and vice versa. In eq 1 the parameters logRB, M_1, M_4, M_7, M_8 have negative coefficients, while Sz, M_2, M_6 and M_9 have positive coefficients. Hence, we can conclude that the lipophilicity decreases with increase in magnitude of logRB, M_1, M_4, M_7 and M_8, while it increases with increase in the magnitude of Sz, M_2, M_6, and M_9. Recall that logRB, M_1, M_4, M_7, M_8 accounts for branching effect, presence of -COOH, mono-substituted anilines, $-CH_2OH$ and thiophene, respectively. It means that such type of molecules present in the set will have negative effect in modelling lipophilic activity of the compounds under present study. Similarly, Sz, M_2, M_6, and M_9 accounts for the presence of ring structure, $-C_6H_5$, $-CONH_2$ and furan ring, respectively, which play positive roles in modelling lipophilicity of the compounds used in the present study. The same is found to be the case with eq 2.

In order to finally confirm our findings we have estimated lipophilicity (logP) using eqs 1, 2 and 3 and compared them with the observed values of lipophilicity. The correlation of observed and estimated lipophilicity are demonstrated in Figs 1–3 giving predictive correlation coefficients (R^2) 0.855, 0.858, 0.859, respectively, for models 8, 10, and 11, thereby finally confirming that the model 11 is the best model for modelling lipophilicity of the compounds used in present case.

We have also used other pairs of topological indices and have also obtained correlations with more than two topological indices but none gave better results than those discussed above.

Several lines of experimental evidences suggest that some lipophilic drugs bind to hydrophobic intra membrane receptor sites via membrane bilayer. Here, the attempted QSAR study may be used for accounting the partitioning of drugs into membranes and the local membrane bilayer environment when the binding event occurs. Our results further show that Szeged index is the most appropriate index in that combination of Sz with one more topological index along with other molecular descriptors (used for various categorries of compounds) will be more useful for the purpose.

Conclusion

From the above mentioned results and discussion we conclude that distance-based topological indices provide structural evidences for modelling lipophilicity. Out of the pool of topological indices used the Szeged index (Sz) is the most appropriate index for this purpose. However, better results are obtained when Sz is combined with one more topological index along with other molecular descriptors (used for various categorries of compounds). The results show that logRB is the most appropriate index to be combined with Sz index for providing topological evidences for modelling lipophilicity of the compounds used.

Experimental

Molecular graphs

The molecular graphs used for the calculation of topological indices W, Sz, χ, B, J, logRB, W* (Table 1) were

carbon–hydrogen as well as hetero-atom hydrogen suppressed graphs.

Lipophilicity (logP). As stated before, the logP were used as reported in the 'star list;' of Hansch et al.[3]

Topological indices

Wiener index (W). The Wiener index (W) is widely used topological index.[26] It is based on the vertex distances of the respective molecular graph.

The molecular graph can be denoted by G and having v_1, v_2, v_3,...,v_n as its vertices. Let $d(v_i,v_j|G)$ stand for the shortest distance between the vertices v_i and v_j. Then the Wiener index is defined as:

$$W = W(G) = 1/2\sum_{i=1}^{n}\sum_{j=1}^{n} d\left(v_i v_j|G\right) \tag{4}$$

Szeged index (Sz). Let e be an edge of the molecular graph G. Let $n_1(e|G)$ be the number of vertices of G lying closer to one end of e; let $n_2(e|G)$ be the number of vertices of G lying closer to the other end of e. Then the Szeged index (Sz) is defined[11,12] as:

$$Sz(G) = Sz = \sum_{e} n_1(e|G)n_2(e|G) \tag{5}$$

with the summation giving over all edges of G.

In cyclic graphs, there are edges equidistant from both the ends of edge e; by definition of Sz such edges are not taken into account.

Balaban index (J). The Balaban index, J (the average distance sum connectivity index) is defined[31,32] by:

$$J = J(G)\frac{M}{\mu+1}\sum_{\text{bonds}}\left(d_i d_j\right)^{-1/2} \tag{6}$$

where M is the number of bonds in a graph G, μ is the cyclomatic number of G and d_i's($i=1,2,3,\ldots,N$) are the distance sums (distance degrees) of atoms in G such that

$$d_i = \sum_{j=1}^{N}(D)_{ij} \tag{7}$$

The cyclomatic number μ of G indicates the number of independent cycles in G and is equal to the minimum number of cuts (removal of bonds) necessary to convert a polycylic structure into an acyclic structure:

$$\mu = M - N + 1 \tag{8}$$

One way to compute the Balaban index (J) for hetero-system is to modified the elements of the distance matrix for hetero-system as follows: (i) The diagonal elements:

$$(D)_{ij} = 1 - (Z_c/Z_i) \tag{9}$$

where $Z_c = 6$ and $Z_i =$ atomic number of the given element.

(ii) The off-diagonal elements:

$$(D)_{ij}d_i = \sum_{r} k_r \tag{10}$$

where the summation is over all bonds. The bond parameter k_r is given by:

$$k_r = 1/b_r\left(Z_c^2/Z_i Z_j\right)$$

where b_r is the bond weight with values: 1 for single bond, 2 for double bond, 1.5 for aromatic bond and 3 for triple bond.

The molecular connectivity indices. The connectivity index $\chi = \chi(G)$ of a graph G is defined by Randic[27,42,43] as under:

$$\chi = \chi(G) = \sum_{ij}\left[\delta_i\delta_j\right]^{-0.5} \tag{11}$$

where δ_i and δ_j are the valence of a vertex i and j, equal to the number of bonds connected to the atoms i and j, in G.

In the case of hetero-systems the connectivity is given in terms of valence delta values δ_i^v and δ_j^v of atoms i and j and is denoted by χ^v. This version of the connectivity index is called the valence connectivity index and is defined[43,44] as under:

$$\chi^v = \chi^v(G) = \sum_{ij}\left[\delta_i^v\delta_j^v\right]^{-0.5} \tag{12}$$

where the sum is taken over all bonds i–j of the molecule. Valence delta values are given by the following expression:

$$\delta_i^v = \frac{Z_i^v - H_i}{Z_i - Z_j - 1} \tag{13}$$

where Z_i is the atomic number of atom i, Z_i^v is the number of valence electron of the atom i and H_i is the number of hydrogen atoms attached to atom i.

Nowadays, the connectivity and the valence connectivity indices expressed by eqs 11 and 12 are termed as first-order connectivity and first-order valence connectivity indices respectively. Lower or higher order indices are also possible which are defined analogously.

Branching index (B) and logRB. The branching index B and RB have been calculated by the method as described by Todeschini et al.[28–30]

***Other molecular descriptors* (M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9).** In order to explain the heterogeneous collection of 140 compounds we have adopted nine molecular descriptors to account for the presence of a particular category of compounds in the set. These molecular descriptors so chosen are: M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9, the details of them are given below.

The molecular descriptors (variables) take on only two values, zero and one. The two values signify that the observation belongs to that particular class of compounds. The numerical values of the molecular descriptors are not intended to reflect a quantitative ordering of categories, but only serve to identify category or class membership. Therefore, they show the significance of a particular class (category) of compounds in a given series (140 compounds used) of drug. They account for the abrupt increase or decrease of a given pharmacological activity at any specific site in the drug molecule. If the coefficient of molecular descriptors carry a negative sign in the regression expression, this makes it very clear that the compounds belonging to this category have considerable lower potency.

The molecular descriptors: M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, and M_9 used for the presence of (1) or absence (0) of –COOH, $-C_6H_5$, monosubstituted phenols, monosubstituted anilines, pyridine nucleous, $-CONH_2$, $-CH_2OH$, thiophene, and furan, respectively.

Regression analysis. Maximum R^2 improvement method to identify prediction models.[35] This method finds the 'best' one variable model, the 'best' two variable model and so forth for the prediction of property/activity. Several models (combinations of variables) were examined to identify combinations of variables with good prediction capabilities. In all regression models developed, a variety of statistics were examined associated with residues, that is, the Wilks-Shapiro test for normality and Cooks D-statistics for outliers, to obtain the most reliable results.[35] Finally, results are discussed on the basis of cross-validation parameters.

Multiple regression analyses for correlating lipophilicity of the present set of compounds with the aforementioned molecular descriptors were carried out using *Regress-1* software as supplied by Professor I. Lukovits, Hungarian Academy of Sciences, Budapest, Hungary. Several multiple regressions were attempted using correlation matrix from this program and the best results were considered and discussed in developing QSAR and hence, for modeling the lipophilicity of the compounds.

Computations. All the computations were carried out in Power Macintosh 9600/233.

Acknowledgements

Authors are thankful to Professor I. Lukovits, Hungarian Academy of Sciences, Budapest, Hungary for providing software to carryout regression analysis and to Prof. Ivan Gutman, Faculty of Science, University of Kragujevac, Yugoslavia for introducing one of the authors (P.V.K.) to this fascinating field of chemical graph theory and topology.

References and Notes

1. Seydal, J. R.; Schaper, K. J. *Chemisch Strukture und BiologicScal Aktivity von Wirkstoflen: Methoden der Quantitativen Struktur Wirkung-Analysis*; Chemie: Weinheim, 1979.
2. Van de Waterbeemeo, H.; Mannhold, R. In Pliska V., Testa, B., Van de Waterbeemeo, H. (Eds.) *Lipophilicity in Drug Action*; VCH: Weinheim, 1996; p 401.
3. Hansch, C.; Leo, A.; Hockman, D. *Exploring QSAR: Hydrolysis and Steric Constants*; American Chemical Society: Washington DC, 1995.
4. Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity*; VCH: Weinheim, 1992.
5. Hansch, C.; Leo, A. *Substitution Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
6. Rekker, R. F.; De Kort, H. M. *Eur. J. Med. Chem.* **1979**, *14*, 479.
7. Schaper, K. J.; Rosado Samitter, M. C. *Quant. Struct. Act. Relat* **1995**, *16*, 224.
8. Khadikar, P. V.; Singh, S.; Shrivastava, A. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1125.
9. Khadikar, P. V.; Phadnis, A.; Shrivastava, A. *Bioorg. Med. Chem.* **2002**, *10*, 1181.
10. Khadikar, P. V.; Mathur, K. C.; Singh, S.; Phadnis, A.; Shrivastava, A.; Mandloi, M. *Bioorg. Med. Chem.* **2002**, *10*, 1761.
11. Gutman, I. *Graph Theory Notes New York* **1994**, *27*, 9.
12. Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 547.
13. In Hungarian, 'Sz' is pronounced as 'S' in English; for instance, the Hungarian spelling of 'September', 'software', 'sex' and 'story' is 'szeptember', 'szoftver', 'szex' and 'sztori' whereas the pronounciation (as well as the meaning) are essentially the same as in English.
14. Mandloi, M.; Sikarwar, A.; Sapre, N. S.; Karmarkar, S.; Khadikar, P. V. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 57.
15. Khadikar, P. V.; Kale, P. P.; Deshpande, N. V.; Karmarkar, S.; Agrawal, V. K. *Comm. Math. Comput. Chem. (MATCH)* **2001**, *43*, 7.
16. Agrawal, V. K.; Shrivastava, R.; Khadikar, P. V. *Bioorg. Med. Chem.* **2001**, *9*, 3287.
17. Agrawal, V. K.; Khadikar, P. V. *Bioorg. Med. Chem.* **2001**, *9*, 2787.
18. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 934.
19. Agrawal, V. K.; Khadikar, P. *V SAR QSAR Environ Res.* **2001**, *12*, 529.

20. Agrawal, V. K.; Joseph, S.; Khadikar, P. V.; Karmarkar, S. *Acta Pharm*. **2000**, *50*, 329.
21. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *Nat. Acad. Sci. Lett*. **2000**, *23*, 165.
22. Mandloi, M.; Agrawal, V. K.; Mathur, K. C.; Karmarkar, S.; Khadikar, P. V. *Acta Pharm* **2000**, *50*, 303.
23. Karmarkar, S.; Joshi, S.; Sharma, V.; Khadikar, P. V. *J. Indian Chem. Soc*. **2000**, *77*, 433.
24. Khadikar, P. V.; Kale, P. P.; Deshpande, N. V.; Agrawal, V. K. *J. Indian Chem. Soc*. **2000**, *77*, 449.
25. Khadikar, P. V.; Joshi, S. *X-ray Spectroscopy* **1995**, *24*, 201.
26. Wiener, H. *J. Am. Chem. Soc*. **1947**, *69*, 17.
27. Randic, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
28. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gorden and Breach: Williston, VT, 2000.
29. Todeschini, R.; Cosonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, 2000.
30. Devillers, J. *Comparative QSAR*; Tailor & Francis: Philadelphia, 1998.
31. Balaban, A. T. *Chem. Phys. Lett*. **1982**, *89*, 399.
32. Khadikar, P. V.; Sharma, S.; Sharma, V.; Joshi, S.; Lukovits, I.; Kaveeshwar, M. *Bull Soc. Chem. Belg* **1997**, *106*, 767.
33. Diudea, M. V., Ed. *QSPR/QSAR Studies by Molecular Descriptors;*. Babes-Bolyai University: Cluj, Romania, 2000.
34. Balaban, A. T. J. *Chem. Inf. Comput. Sci*. **1992**, *32*, 23.
35. Chaterjee, S.; Hadi, A. S.; Price, B. *Regression Analysis by Examples*, 3rd ed.; Wiley: New York, 2000.
36. Mandloi, M.; Sikarwar, A.; Sapre, N.; Karmarkar, S.; Khadikar, P. V. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 57.
37. Khadikar, P. V.; Karmarkar, S.; Agrawal, V. K. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 934.
38. Pogliani, L. *Amino Acids* **1994**, *6*, 141.
39. Agrawal, V. K.; Srivastava, R.; Khadikar, P. V. *Acta Pharm*. **2001**, *51*, 117.
40. Agrawal, V. K.; Sharma, R.; Khadikar, P. V. *Bioorg. Med. Chem*. **2002**, *10*, 1361.
41. Kauffan, G.W., Jurs, P.C. J. Chem. Inf. Comput. Sci. (In press).
42. Randic, M. *Croat. Chem. Acta* **1993**, *66*, 289.
43. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Relationship*; Wiley: New York, 1986.
44. Kier, L. B.; Hall, L. H. *Molecular Structure Description*; Academic: New York, 1999.