Inter*net* Electronic Journal of
Molecular Design

May 2003, Volume 2, Number 5, Pages 334–347

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday
Part 9

Guest Editor: Jun–ichi Aihara

QSPR Modeling of Lipophilicity by Means of Correlation Weights of Local Graph Invariants

Pablo J. Peruzzo,[1] Damián J. G. Marino,[1] Eduardo A. Castro,[2] and Andrey A. Toropov[3]

[1] CIMA, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Calles 47 y 115, La Plata 1900, Argentina
[2] CEQUINOR, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 962, La Plata 1900, Argentina
[3] Vostok Innovation Company, Azimstreet 4, Tashkent 700047, Uzbekistan

Received: February 27, 2002; Revised: September 3, 2002; Accepted: March 17, 2002; Published: May 31, 2003

Citation of the article:
P. J. Peruzzo, D. J. G. Marino, E. A. Castro, and A. A. Toropov, QSPR Modeling of Lipophilicity by Means of Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des*. **2003**, *2*, 334–347, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

QSPR Modeling of Lipophilicity by Means of Correlation Weights of Local Graph Invariants[#]

Pablo J. Peruzzo,[1] Damián J. G. Marino,[1] Eduardo A. Castro,[2,]* and Andrey A. Toropov[3]

[1] CIMA, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Calles 47 y 115, La Plata 1900, Argentina
[2] CEQUINOR, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, C.C. 962, La Plata 1900, Argentina
[3] Vostok Innovation Company, Azimstreet 4, Tashkent 700047, Uzbekistan

Received: February 27, 2002; Revised: September 3, 2002; Accepted: March 17, 2002; Published: May 31, 2003

Internet Electron. J. Mol. Des. **2003**, *2* (5), 334–347

Abstract

A quantitative structure–property modeling of the log *P* (octanol/water partition coefficient) of 76 industrial chemicals is presented. Estimations are performed by means of correlation weighting of local invariants of labeled hydrogen–filled graphs. Results are quite satisfactory, with lower average deviations than other calculations performed with similar theoretical methods. Some possible applications and further extensions of the computation procedure to estimate other physico–chemical or biological properties are mentioned.

Keywords. QSPR; quantitative structure–property relationships; topological indices; lipophilicity; correlation weights; local graph invariants; octanol/water partition coefficient.

1 INTRODUCTION

Lipophilicity is a measure of the degree to which a given molecule prefers hydrophobic nonpolar environments to water. The most common experimental measure of lipophilicity is the logarithm of the partition coefficient for a solute distributing itself between water and some organic solvent, such as 1–octanol or chloroform. This quantity is abbreviated as $\log P$ and has been measured experimentally for a wide range of organic compounds [1]:

$$\log P = \log\left([S]_{org} / [S]_{aq}\right) \tag{1}$$

The partition coefficient for octanol–water ($\log P_{\text{ow}}$) has become the preferred measure for lipophilicity in the development of biologically active molecules, in which transport across

[#] Dedicated to Professor Haruo Hosoya on the occasion of the 65th birthday.
* Correspondence author; phone: 054–221–4256742; fax: 054–221–4259485; E–mail: castro@quimica.unlp.edu.ar.

biological membranes is often critical [2–5]. Since the concept of lipophilicity is so important in chemistry, many schemes have been developed to estimate this property as expressed by the partition coefficient log *P*. Some of the best known methods, such as Leo's CLOGP program [6], rely upon summing group contributions of structural fragments to calculate log *P* directly from the two–dimensional structure of a molecule. Methods for calculating log *P* were surveyed in a 1997 review [2]. The fragment–based methods are reasonably accurate and very fast, but they suffer from a few limitations, such as the need for many parameters and the inability to calculate log *P* for structures containing completely novel structural fragments.

The possibility to calculate log *P* directly using explicit solvent or continuum–based simulations opens up many new opportunities for modeling chemical properties related to lipophilicity. Both the explicit solvent and continuum calculations provide significantly more structural detail than the fragment–based methods that are more commonly used today. Furthermore, even when speed is critical, such as with large corporate databases, the more general approaches to log *P* can be used to derive missing parameters for fragment–based methods. Thus, direct calculation of log *P* could become very important in ensuring complete coverage of corporate structural databases, virtual compound libraries, or collections of acquisition compounds.

Since the work of Meyer and Overton a century ago [7,8], lipophilicity has been recognized as a meaningful parameter in structure–activity relationship studies, and which the epoch–making contributions of Hansch [9] has become the single most informative and successful physicochemical property in medicinal chemistry [3,10,11]. Not only has lipophilicity found innumerable applications in quantitative structure–activity and structure–disposition relationships, but its study has revealed a wealth of information on molecular structure.

The importance of calculated log *P* is also enhanced by the rapid development of combinatorial chemistry. Computational methods are indeed the only techniques allowing a realistic of the lipophilicity of molecular fragments linked to inert supports. Moreover, the number of lead compounds generated by combinatorial chemistry calls for more accurate methods able to optimize and select these drug candidates. In this context, Quantitative Structure Activity (Property) Relationships (QSAR/ QSPR) techniques based on calculated log *P* offer tools to assess both solvation and entropy effects, simplifying the estimate of the binding free energy of ligands [12–15]. The main advantage of this sort of methodology is its independence with respect to Molecular Orbital (MO) theory, thus avoiding a rather troublesome, time–consuming process.

The aim of this paper is to describe a QSPR modeling of log *P* (octanol/water) of a diverse set of 76 industrial chemicals by means of correlation weighting of local invariants of Labeled Hydrogen–Filled Graphs (LHFGs). This method was proposed recently [16–19] and it has proved to be quite useful to predict several physical chemistry properties [20–26].

This paper is organized as follows: Section 2 deals with the basic definitions related to the topological descriptor, giving their foundatons and pointing out its usefulness as well as analyzing the antecedents of this particular topological index. Section 3 describes the methodology applied to obtain the regression equations. Section 4 presents some illustrative numerical results, comparing them with other data arising from alternative theoretical procedures. Finally, in Section 5 we discuss the possibility of extending the use of this sort of molecular descriptor to a different set of molecules and for studying other physical chemistry properties and biological activities.

2 MATERIALS AND METHODS

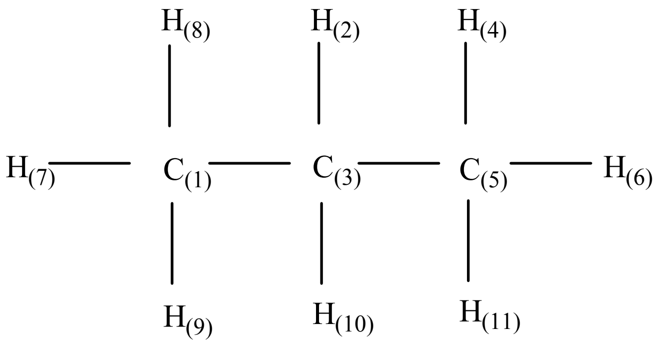
2.1 Correlation Weighting of Local Graph Invariants

The last three decades witnessed an upsurge of interest in applications of graph theory in chemistry [27–31]. Graph theory is a basic tool for fostering alternative ways to solve chemical problems, both by the high degree of abstraction evidenced by the generality of such concepts as points, lines and neighborhoods as well as by the combinatorial derivation of many graph–theoretical concepts which correspond to the essence of chemistry considered as “the study of combination between atoms” [32]. This method offers a wide variety of concepts and procedures of significant importance to chemistry.

A graph G is defined as a finite non–empty set $V(G)$ of N–vertices (points) together with a set $E(G)$ of edges (lines), the latter being unordered pairs of distinct vertices. Then, by definition, every ghraph is finite and has no loops (in edge initiating from and ending in one and the same vertex) and multiple edges. When two vertices x and y are joined by an edge $e = \{x, y\}$, vertices x and y are said to be adjacent and each of them is incident with the edge e. As a matter of fact, the structural (constitutional) formula of a chemical compound may be regarded as a molecular graph (MG), where the vertices represent atoms while the edges stand for valence bonds. The graph–theoretical characterization of molecular structure is most often made by its translation into molecular descriptors, such as topological indices.

A topological index is a real number, associated in an arbitrary way, characterizing the graph. It is based on a certain topological feature of the corresponding MG and represents a graph invariant, that is to say, it does not depend on the vertex numbering [33]. The main field of application of topological indices is the structure–property and structure–activity quantitative correlations. Different graph characteristics or invariants have been used in the definition of molecular topological indices. For example, the kind of chemical element, the vertex degree $^{\circ}\chi(i)$ and the Morgan vertex degrees o first–order $^{1}\chi(i)$ are well–known typical local invariants [34,35].

Values $^{\circ}\chi(i)$ and $^{1}\chi(i)$ can be computed from the adjacency matrix. For example, the adjacency matrix for propane is presented below:



$$\mathbf{A} = \{a_{ij}\} = \begin{array}{c|ccccccccccc}
 & C_{(1)} & H_{(2)} & C_{(3)} & H_{(4)} & C_{(5)} & H_{(6)} & H_{(7)} & H_{(8)} & H_{(9)} & H_{(10)} & H_{(11)} \\
C_{(1)} & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
H_{(2)} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_{(3)} & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
H_{(4)} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
C_{(5)} & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
H_{(6)} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
H_{(7)} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H_{(8)} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H_{(9)} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H_{(10)} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
H_{(11)} & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}$$

together with the corresponding °χ(i) and $^1\chi$(i) values:

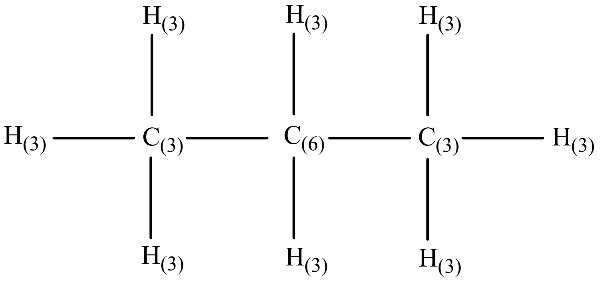
Atom	$C_{(1)}$	$H_{(2)}$	$C_{(3)}$	$H_{(4)}$	$C_{(5)}$	$H_{(6)}$	$H_{(7)}$	$H_{(8)}$	$H_{(9)}$	$H_{(10)}$	$H_{(11)}$
°χ(i)	4	1	4	1	4	1	1	1	1	1	1
$^1\chi$(i)	7	4	10	4	7	4	4	4	4	4	11

which are derived from the definitions:

$$^0\chi(i) = \sum_{j=1}^{n} a_{ij} \tag{2}$$

$$^1\chi(i) = \sum_{edges\{i,j\}} {}^0\chi(i) \tag{3}$$

The new topological index represents molecular structures via values of the correlation weights of local invariants of LHFGs. As local invariants we use numbers of paths of length 2, which have been suggested three years ago by Randić [36]. The local invariants will be denoted as P2$_k$. Values of the local invariants on vertices of LHFG for propane are:



2.2 Computational Methodology

We have chosen a diverse set of 76 industrial chemicals provided in CLOGP [37]. This set includes molecules containing C, H, O, Cl, S, F, Br and N atoms and it is rather representative because there are alcohols, nitriles, ketones, amines, esters, furans, amides, ethers, aldehydes, nitro and halogen derivatives and hydrocarbons. Table 1 lists the compounds used in this study and the value of log P_{ow} for each compound together with previous theoretical results. This particular choice is identical to that used by Basak and Grunwald [38] and was selected in order to perform a direct comparison of the present method with another procedure based on molecular structural similarity.

Table 1. List of 76 molecules analysed with their measured log *P* values and some estimated data using K–nearest neighbor calculation by similarity methods [38]

Molecule	log P (exp)	log P (theor) 1	[AP method] 2	log P (theor) 1	[ED method] 2
Methanol *	–0.77	–	–	–0.31	–0.22
Acetonitrile *	–0.34	0.16	0.16	0.16	–0.08
Ethanol °	–0.31	0.25	0.06	–0.13	0.06
Acetone *	–0.24	0.29	0.23	0.05	0.53
Ethylamine °	–0.13	0.48	0.08	–0.31	–0.03
2–Propanol °	0.05	0.35	0.55	–0.24	0.26
Propionitrile °	0.16	–0.31	–0.22	0.25	0.36
Methyl acetate *	0.18	0.29	0.51	0.73	1.02
1–Propanol °	0.25	0.88	0.75	0.48	0.68
2–Butanone °	0.29	0.18	0.55	0.91	0.76
2–Methyl–2–propanol *	0.35	0.05	–0.13	0.05	–0.10
Tetrahydrofuran *	0.46	3.44	3.33	0.58	0.73
Propylamine *	0.48	0.97	0.61	0.25	0.61
Diethylamine *	0.58	0.89	0.89	0.89	0.68
2–Butanol °	0.61	0.25	0.51	0.29	0.60
Benzamide *	0.64	0.90	1.19	1.85	2.15
Pyridine *	0.65	2.13	1.51	0.90	1.19
Ethyl acetate °	0.73	0.91	1.11	0.18	0.75
2–Methyl–1–propanol *	0.76	0.61	0.33	0.29	0.60
Cyclohexanone *	0.81	3.44	2.71	2.99	2.86
1–Butanol °	0.88	1.56	0.91	0.90	1.23
Diethyl ether °	0.89	0.58	0.66	0.58	0.52
Aniline *	0.90	2.99	2.86	1.48	1.71
2–Pentanone °	0.91	0.73	1.05	0.29	0.83
Butylamine °	0.97	1.49	0.98	0.88	0.68

Table 1. (Continued)

Molecule	log P (exp)	log P (theor) 1	[AP method] 2	log P (theor) 1	[ED method] 2
N,N–Dimethylformamide *	1.01	1.38	2.05	0.18	0.45
4–Fluoroaniline *	1.15	1.39	1.14	1.85	1.25
Ethyl acrylate *	1.32	0.73	1.05	0.73	0.45
Methyl methacrylate *	1.38	0.18	0.75	1.32	0.75
2–Hexanone °	1.38	1.98	1.44	0.91	1.20
4–Toluidine °	1.39	3.15	2.54	1.94	1.42
Benzaldehyde *	1.48	0.64	1.25	0.90	1.42
1,2–Dichloroethane *	1.48	0.25	0.36	–0.34	–0.29
Amylamine °	1.49	2.06	1.52	1.56	1.22
Isopropyl ether *	1.52	0.05	0.47	3.15	2.02
1–Pentanol °	1.56	2.03	1.45	1.49	1.18
Nitrobenzene °	1.85	2.45	1.96	0.64	0.90
Hexanoic acid *	1.92	1.98	2.00	1.98	2.02
4–Methylphenol *	1.94	3.15	2.27	1.39	1.14
2–Heptanone *	1.98	1.92	1.65	2.06	2.04
1–Hexanol °	2.03	2.72	2.14	2.06	2.02
Hexylamine °	2.06	2.57	2.03	2.03	2.00
Benzene *	2.13	0.65	1.06	3.44	1.95
1,1,2,2–Tetrachloroethane*	2.39	3.40	2.91	3.40	3.52
Trichloroethylene °	2.42	3.40	3.39	0.16	1.88
m–Nitrotoluene °	2.45	1.85	2.53	0.64	0.90
1,1,1–Trichloroethane *	2.49	2.83	2.61	0.35	1.59
n–Heptylamine °	2.57	2.06	2.39	2.72	2.37
Ethyl benzoate °	2.64	0.64	2.45	0.64	1.55
1–Heptanol °	2.72	2.97	2.50	2.57	2.77
Toluene °	2.73	3.15	2.02	3.15	1.98
Tripropylamine *	2.79	2.97	3.09	3.21	3.26
Carbon tetrachloride *	2.83	2.49	2.44	2.49	1.42
1–Naphthol °	2.84	3.30	3.69	4.26	4.09
1–Octanol °	2.97	2.72	2.72	2.72	2.64
Bromobenzene °	2.99	0.90	1.81	0.81	1.14
o–Xylene °	3.12	3.78	3.25	3.66	3.44
p–Xylene °	3.15	3.78	2.86	3.20	1.83
Ethylbenzene *	3.15	2.73	3.20	2.73	2.05
m–Xylene °	3.20	3.78	3.25	3.15	3.41
Butyl ether *	3.21	2.97	2.77	2.97	2.77
Naphthalene °	3.30	4.09	3.46	4.09	3.70
N,n–Diethylaniline *	3.31	4.26	3.70	3.15	3.15
1,2–Dichlorobenzene °	3.38	4.02	3.81	3.64	2.14
Tetrachloroethylene *	3.40	2.42	2.41	2.39	3.02
Cyclohexane *	3.44	2.57	2.64	2.13	1.47
1,3–Dichlorobenzene °	3.60	4.02	3.70	2.99	3.18
1,2–Dibromobenzene *	3.64	2.99	3.06	3.38	3.70
Isopropylbenzene °	3.66	3.15	2.94	3.12	3.45
1,2,4–Trimethylbenzene °	3.78	3.15	3.13	3.66	3.43
Acenaphthene *	3.92	2.84	3.07	2.84	3.46
1,2,4–Trichlorobenzene °	4.02	4.82	4.10	3.64	3.64
Biphenyl °	4.09	3.30	3.07	3.30	3.07
Butylbenzene *	4.26	3.31	3.23	1.98	2.02
1,2,4,5–Tetracholobenzene*	4.82	4.02	4.59	3.64	3.51
Pentachlorobenzene *	5.17	4.82	4.42	3.83	3.64
Average absolute deviation	–	0.72	0.63	0.64	0.60

The QSPR modeling for log *P* has been made by means of the following calculation scheme [17,18,39,40]:

1. A computer program read adjacency matrices corresponding to the different MGs.

2. These matrices are translated into adjacency matrices associated with the AOMGs.

3. Local invariants are then computed.

4. A suitable optimization procedure allows one to find the correlations weights (CWs) of the local invariants which yield maximum values for the correlation coefficient in the regression equations for log *P* via descriptors DCW, defined as:

$$DCW = \sum_{i=j}^{n} \left[CW\{a_i\} \right] + CW\{P2_i\} \tag{4}$$

where *n* is the number of vertices in the LHFG, CW{a_i} is the correlation weight associated with the pressence of vertex a_i, such as H, C, N, O, F, Cl, and Br, and CW (P2_i) is the correlation weight of the number of order two–paths starting from the i^{th} vertex in the molecular graph. Table 2 shows results of three probes using this procedure.

Table 2. Three probes for local invariants

Atom	CW(a_i) – Probe 1	CW(a_i) – Probe 2	CW(a_i) – Probe 3
H	0.283	0.252	0.394
C	1.099	1.105	0.978
N	0.028	0.017	0.024
O	0.028	0.017	0.024
F	0.028	0.017	0.024
Cl	5.481	5.178	5.313
Br	7.807	7.342	7.525
P2_i	CW(P2_i) – Probe 1	CW(P2_i) – Probe 2	CW(P2_i) – Probe 3
0000	1.485	1.446	1.440
0001	1.087	1.005	1.034
0002	0.404	0.375	0.352
0003	0.223	0.202	0.132
0004	0.760	0.687	0.721
0005	0.028	0.017	0.024
0006	1.460	1.337	1.349
0007	3.238	2.979	3.141
0008	4.440	4.122	4.157
0009	0.375	0.340	0.227
0010	0.028	0.017	0.024

An illustrative example of descriptor calculation for acetone is given in Table3.

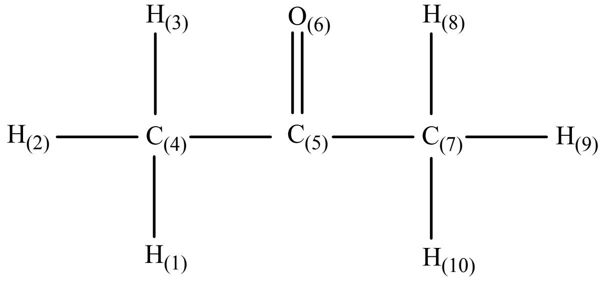


Table 3. Calculation of acetone descriptor

Atom type	Numbering	CW(a_i)	P2$_i$	CW(P2$_0$)
H	1	0.283	3	0.223
H	2	0.283	3	0.223
H	3	0.283	3	0.223
C	4	1.099	2	0.404
C	5	1.099	6	1.460
O	6	0.028	2	0.404
C	7	1.099	2	0.404
H	8	0.283	3	0.223
H	9	0.283	3	0.223
H	10	0.283	3	0.223
ΣCW	–	5.023	–	4.010
DCW		9.033		

These values allow us to determine CWs, which are then applied in a least–square fitting method to obtain log P via a general relationship:

$$\log P = a + b(DCW)^r + c(DCW)^s + d(DCW)^t \tag{5}$$

where *r*, *s*, *t* are rational numbers and coefficients *a*, *b*, *c*, and *d* are determined by regression analysis.

3 RESULTS AND DISCUSSION

The original set of 76 molecules was split up into two subsets: a working set comprising 38 molecules and a test set with the remaining 38 molecules. During the optimization procedure we experimented with three probes. Different choices for each set were made, but final results are nearly independent of them. The most significant statistical results are given in Table 4.

Table 4. Statistical results for log *P*

$$\log P = a + b \times (DCW) \tag{6}$$

Probe	Training set					Test set		
	a	*b*	*r*	*s*	F	*r*	*s*	F
1	0.166	–1.381	0.9411	0.500	279	0.9591	0.392	414
2	0.175	–1.366	0.9412	0.500	280	0.9618	0.382	444
3	0.174	–1.380	0.9416	0.498	282	0.9596	0.388	418

$$\log P = a + b \times (DCW) + c \times (DCW)^2 \tag{7}$$

Probe	Training set						Test set		
	a	*b*	*c*	*r*	*s*	F	*r*	*s*	F
1	–1.5168	0.1804	–3.44E–4	0.9413	0.514	280	0.9586	0.378	408
2	–1.5202	0.1924	–4.35E–4	0.9415	0.513	281	0.9612	0.367	437
3	–1.5198	0.1902	–3.91E–4	0.9418	0.511	282	0.9590	0.375	412

Table 4. (Continued)

$$\log P = a + b \times (DCW) + c \times (DCW)^2 + d \times (DCW)^3 \tag{8}$$

Probe	Training set							Test set		
	a	*b*	*c*	*d*	*r*	*s*	F	*r*	*s*	F
1	–1.408	0.1614	6.0690E–4	–1.3940E–5	0.9413	0.521	280	0.9590	0.375	412
2	–1.392	0.1685	8.2086E–4	–1.9447E–5	0.9416	0.520	281	0.9619	0.356	446
3	–1.404	0.1689	7.2568E–4	–1.7206E–5	0.9419	0.519	283	0.9594	0.373	416

$$\log P = a + b \times (DCW)^{1/2} \tag{9}$$

Probe	Training set					Test set		
	a	*b*	*r*	*s*	F	*r*	*s*	F
1	–4.3480	1.4366	0.9355	0.523	252	0.9509	0.402	340
2	–4.3221	1.4725	0.9358	0.522	254	0.9536	0.394	361
3	–4.3499	1.4740	0.9361	0.521	255	0.9514	0.400	343

$$\log P = a + b \times (DCW)^{1/3} \tag{10}$$

Probe	Training set					Test set		
	a	*b*	*r*	*s*	F	*r*	*s*	F
1	–7.2487	3.4526	0.9295	0.546	229	0.9459	0.421	306
2	–7.2902	3.5046	0.9299	0.545	230	0.9486	0.411	323
3	–7.2537	3.5135	0.9302	0.543	231	0.9464	0.418	309

Table 5. Estimated log *P* values using correlation weigths of local graph invariants and experimental data

Molecule	log *P* (exp)	log *P* (°)	Residual	log *P* Eq. (5)	Residual	log *P* Eq. (8)	Residual
Methanol	–0.77	–0.62	–0.15	–0.50	0.35	–1.05	1.4
Acetonitrile	–0.34	–0.36	0.02	–0.35	0.37	–0.78	1.15
Ethanol	–0.31	–0.21	–0.1	–0.17	0.07	–0.47	0.54
Acetone	–0.24	–0.21	–0.03	0.12	–0.15	–0.02	–0.13
Ethylamine	–0.13	0.03	–0.16	–0.29	0.13	–0.66	0.79
2–Propanol	0.05	0.23	–0.18	0.61	–0.79	0.64	–1.43
Propionitrile	0.16	0.30	–0.14	–0.02	–0.12	–0.24	0.12
Methyl acetate	0.18	0.44	–0.26	0.12	–0.38	–0.02	–0.36
1–Propanol	0.25	0.43	–0.18	0.41	–0.59	0.38	–0.97
2–Butanone	0.29	0.42	–0.13	0.44	–0.57	0.42	–0.99
2–Methyl–2–propanol	0.35	0.34	0.01	0.48	–0.47	0.47	–0.94
Tetrahydrofuran	0.46	0.51	–0.05	0.98	–1.03	1.08	–2.11
Propylamine	0.48	0.69	–0.21	0.30	–0.51	0.22	–0.73
Diethylamine	0.58	0.61	–0.03	0.62	–0.65	0.65	–1.3
2–Butanol	0.61	0.85	–0.24	1.20	–1.44	1.32	–2.76
Benzamide	0.64	0.71	–0.07	1.84	–1.91	1.99	–3.9
Pyridine	0.65	0.68	–0.03	0.70	–0.73	0.74	–1.47
Ethyl acetate	0.73	0.78	–0.05	0.45	–0.5	0.43	–0.93
2–Methyl–1–propanol	0.76	0.84	–0.08	0.63	–0.71	0.65	–1.36
Cyclohexanone	0.81	1.02	–0.21	1.59	–1.8	1.74	–3.54
1–Butanol	0.88	1.07	–0.19	1.00	–1.19	1.10	–2.29
Diethyl ether	0.89	1.04	–0.15	0.74	–0.89	0.80	–1.69
Aniline	0.90	1.21	–0.31	1.46	–1.77	1.60	–3.37
2–Pentanone	0.91	1.01	–0.1	0.62	–0.72	0.64	–1.36
Butylamine	0.97	1.20	–0.23	0.88	–1.11	0.96	–2.07
n.n–Dimethylformamide	1.01	1.01	0	0.78	–0.78	0.85	–1.63
4–Fluoroaniline	1.15	1.21	–0.06	1.42	–1.48	1.56	–3.04
Ethyl acrilate	1.32	1.31	0.01	0.77	–0.76	0.83	–1.59
Methyl methacrylate	1.38	1.75	–0.37	0.90	–1.27	0.99	–2.26

Table 5. (Continued)

Molecule	log *P* (exp)	log *P* (°)	Residual	log *P* Eq. (5)	Residual	log *P* Eq. (8)	Residual
2–Hexanone	1.38	1.39	–0.01	1.61	–1.62	1.76	–3.38
4–Toluidine	1.39	1.70	–0.31	2.24	–2.55	2.38	–4.93
Benzaldehyde	1.48	1.76	–0.28	1.61	–1.89	1.76	–3.65
1.2–Dichloroethane	1.48	2.12	–0.64	1.28	–1.92	1.42	–3.34
Amylamine	1.49	1.65	–0.16	1.98	–2.14	2.13	–4.27
Isopropyl ether	1.52	1.58	–0.06	2.32	–2.38	2.44	–4.82
1–Pentanol	1.56	1.58	–0.02	1.58	–1.6	1.73	–3.33
Nitrobenzene	1.85	1.21	0.64	1.38	–0.74	1.51	–2.25
Hexanoic acid	1.92	1.74	0.18	1.88	–1.7	2.03	–3.73
4–Methylphenol	1.94	2.09	–0.15	2.01	–2.16	2.16	–4.32
2–Heptanone	1.98	1.80	0.18	2.20	–2.02	2.33	–4.35
1–Hexanol	2.03	1.91	0.12	2.17	–2.05	2.30	–4.35
Hexylamine	2.06	2.01	0.05	2.05	–2	2.19	–4.19
Benzene	2.13	2.24	–0.11	1.17	–1.28	1.29	–2.57
1.1.2.2–Tetrachloroethane	2.39	2.90	–0.51	3.00	–3.51	3.05	–6.56
Trichlroethylene	2.42	2.73	–0.31	2.17	–2.48	2.30	–4.78
m–Nitrotoluene	2.45	1.70	0.75	2.16	–1.41	2.30	–3.71
1,1,1–Trichloroethane	2.49	2.41	0.08	2.14	–2.06	2.28	–4.34
n–Heptylamine	2.57	2.40	0.17	2.63	–2.46	2.73	–5.19
Ethyl benzoate	2.64	2.20	0.44	2.41	–1.97	2.53	–4.5
1–Heptanol	2.72	2.30	0.42	2.75	–2.33	2.83	–5.16
Toluene	2.73	2.64	0.09	1.95	–1.86	2.10	–3.96
Tripropylamine	2.79	3.22	–0.43	2.91	–3.34	2.97	–6.31
Carbon tetrachloride	2.83	2.60	0.23	2.83	–2.6	2.90	–5.5
1–Naphthol	2.84	2.72	0.12	2.92	–2.8	2.98	–5.78
1–Octanol	2.97	2.88	0.09	3.34	–3.25	3.32	–6.57
Bromobenzene	2.99	2.87	0.12	2.40	–2.28	2.52	–4.8
o–Xylene	3.12	3.08	0.04	2.74	–2.7	2.82	–5.52
p–Xylene	3.15	3.08	0.07	2.74	–2.67	2.82	–5.49
Ethylbenzene	3.15	3.08	0.07	2.28	–2.21	2.41	–4.62
m–Xylene	3.20	3.08	0.12	2.74	–2.62	2.82	–5.44
Butyl ether	3.21	2.84	0.37	3.08	–2.71	3.11	–5.82
Naphthalene	3.30	3.32	–0.02	2.86	–2.88	2.93	–5.81
n,*n*–Diethylaniline	3.31	3.09	0.22	3.53	–3.31	3.48	–6.79
1.2–Dichlorobenzene	3.38	3.02	0.36	2.89	–2.53	2.95	–5.48
Tetrachloroethylene	3.40	3.30	0.1	3.03	–2.93	3.07	–6
Cyclohexane	3.44	2.85	0.59	2.14	–1.55	2.28	–3.83
1,3–Dichlorobenzene	3.60	3.02	0.58	2.89	–2.31	2.95	–5.26
1,2–Dibromobenzene	3.64	3.68	–0.04	3.64	–3.68	3.57	–7.25
Isopropylbenzene	3.66	3.65	0.01	3.38	–3.37	3.36	–6.73
1,2,4–Trimethylbenzene	3.78	3.60	0.18	3.53	–3.35	3.47	–6.82
Acenaphthene	3.92	4.02	–0.1	4.15	–4.25	3.96	–8.21
1,2,4–Trichlorobenzene	4.02	3.61	0.41	3.75	–3.34	3.65	–6.99
Biphenyl	4.09	2.00	2.09	3.71	–1.62	3.62	–5.24
Butylbenzene	4.26	4.24	0.02	3.45	–3.43	3.41	–6.84
1,2,4,5–Tetrachlorobenzene	4.82	2.02	2.8	4.60	–1.8	4.29	–6.09
Pentachlorobenzene	5.17	1.41	3.76	5.46	–1.7	4.89	–6.59
Average absolute deviation (*)	–	0.34	–	0.36	–	0.40	–
Average absolute deviation (**)	–	0.25	–	0.31	–	0.34	–
Average absolute deviation (***)	–	0.29	–	0.34	–	0.37	–

(°) *Doklady Academii Nauk* **2000**, *374*, 786.

(*) Training set

(**) Test set

(***) Complete set

Finally, in Table 5 we display some theoretical results derived from the present calculation procedure together with experimental data and Pasyukov *et al.*'s theoretical results [41]. We have inserted these theoretical results here since they were derived on the basis of an optimal selection of the measure of molecular similarity.

Since our molecular set is identical to that employed by Basak and Grunwald [38] and Pasyukov *et al.* [41], it allows us to make a direct comparison with previous theoretical results. The analysis of the reported numerical data shows clearly that present predictions of log *P* (octanol/water) are better than previous ones derived from two molecular similarity measures based on atom pairs and topological indices. In fact, when comparing average absolute deviations, our results are better by approximately a factor of two.

Another way to make direct comparisons is illustrated in Table 6, where we report regression coefficients for different methods. In order to judge present results, one must take into account the numerical data reported for molecules belonging to the tests set are real predictions, since coefficients for regression equations were determined from data corresponding to the test set, while such discrimination among molecules were not made in the paper reported by Basak and Grunwald [38].

Regarding Pasyukov *et al.*'s results, statistical parameters are nearly similar. However, these authors have not split up the molecular set into a training set and a test set so that they have not made real predictions. Notwithstanding this drawback, it must be taken into account their results are quite good and probably this minor detail does not diminish the fact their method for predicting the properties of compounds is the most accurate one compared to other molecular similarity measure procedures, as stated explicitly by the authors [41].

Table 6. Regression coefficients for different log *P* estimations

AP method [38]		ED method [38]		Present calculation - probe 3	
K	*r*	K	*r*	Eq.	*r*
1	0.774	1	0.788	5	0.9416
2	0.854	2	0.821	6	0.9418
3	0.869	3	0.814	7	0.9419
4	0.874	4	0.846	8	0.9361
5	0.854	5	0.845	9	0.9362

Predictions do not change significantly when resorting to different algebraic forms of the quantitative relationships (*i.e.* Eqs. (5)–(9)). Besides, polynomial relationships do not improve very much when increasing the algebraic orders (compare, for example, results derived from Eqs. (5) and (6)). Regarding the different probes chosen in this work, they yield nearly the same reuslts, although probe 3 is slightly better than the other two.

4 CONCLUSIONS

Numerical data presented in the previous section make clear the rather good quality results based on the correlation weighting of local invariants of AOMGs, which yield very accurate log *P* (octanol/water) and numerical correlations with significantly low standard errors. The comparison with other theoretical predictions derived from molecular similarity measures based on atom pairs and topological indices for log *P* is favorable for our present approach. This kind of flexible descriptor does not present any difficulty to implement the corresponding numerical algorithm and besides these results are in line with some previous ones [16–26].

These results point out the possibility to extend this sort of method for other molecules and/or physical chemistry properties and biological activities resorting to this new topological descriptor. It also could be interesting to employ multiple regression analysis supported by topological descriptors and molecular indices combined with the orthogonalization procedure in order to obtain optimum QSR and QSPR models that most probably will lead to a meaningful interpretation of the regression formulae. Currently, work along these lines is being carried out in our laboratories and results will be presented elsewhere in the forthcoming future.

We deem suitable to make a final comment on the definition of the Oxc descriptor. In fact, we have employed an additive relationship between $\mathrm{CW}\{a_i\}$ and $\mathrm{CW}(\mathrm{P2}_i)$, *i.e.* Eq. (4), but it should be equally valid to employ other sort of algebraic relationships between local invariants and elements of the adjacency matrix. This possibility has been explored before and results were quite satisfactory [16–18,39], so that they represent an interesting chance to extend this sort of study in QSAR and QSPR models.

Acknowledgment

The authors thank very much some referee's suggestions and comments that have been helpful to improve the final version of this paper.

5 REFERENCES

[1] C. H. Reynolds and S. A. Best, A Fast Molecular Simulation to Calculate Lipophilicity, *CHEMTECH*, November **1998**, 28–34.

[2] P. Carrupt, B. Testa and P. Gaillard, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., Wiley–VCH, New York, 1997; 1899 Vol. *11*, pp.241–345.

[3] C. Hansch and A. J. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley, New York, 1979.

[4] C. Hansch and A. Leo, *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; Americal Chemical Society, Washington, D.D., 1995.

[5] Y. C. Martin, *Quantitative Drug Design: A Critical Introduction*, Marcel Dekker, New York, 1978.

[6] A. Leo, *CLOGP*, Daylight Chemical Information Systems, Mission Viejo, Ca.

[7] H. Meyer, Zur Theorie der Alkoholnarkose. I. Welche Eigenschaft der Anaesthetika bedingt ihre narkotische Wirkung?, *Arch. Exp. Pathol. Pharmakol.*, **1899**, *42*, 109.

[8] E. Overton, Über die allgemeinen osmotischen Eigenschaften der Zelle, ihre vermutlichen Ursachen und ihre Bedeutung für die Physiologie, *Vierteljahrsschr. Naturforsch. Ges. Zürich* **1899**, *44*, 87.

[9] C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, Biological Actibity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients, *Nature*. **1962**, *194*, 178.

[10] F. Helmer, K. Kiehs and C. Hansch, The Linear Free–Energy Relation Between Partition Coefficients and the Binding, and Conformational Perturbation of Macromolecules by Small Organic Compounds, *Biochemistry*, **1968**, *7*, 2858.

[11] R. F. Rekker, *The Hydrophobic Fragmental Constant*, W. T. Nauta and R. F. Rekker, Eds., Elsevier, Amsterdam, 1977.

[12] T. P. Lybrand, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., Wiley–VCH, New York, 1990, Vol. *1*, pp.295–320.

[13] T. P. Straatsma, in *Reviews in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., Wiley–VCH, New York, 1996, Vol. *9*, pp.81–127.

[14] M. S. Tute, in *Advances in Drug Research*, B. Testa and U. A. Meyer, Eds., Academic Press, London, 1995, Vol. *26*, pp. 45–142.

[15] L. M. Balbes, S. W. Mascarella and D. B. Boyd, in *Review in Computational Chemistry*, K. B. Lipkowitz and D. B. Boyd, Eds., Wiley–VCH, New York, 1994, Vol. *5*, pp. 337–379.

[16] A. A. Toropov, N. L. Voropaeva, I. N. Ruban and S. Sh. Rashidova, Quantitative Structure–Property Relationships for Binary Polymer–Solvent Systems: Correlation Weighting of the Local Invariants of Molecular Graphs, *Polymer Sci. Ser. A*. **1999**, *41*, 975–985.

[17] A. A. Toropov and A. P. Toropova, Prediction of Heteroaromatic Amine Mutagenicity by Means of Correlation Weighting of Atomic Orbital Graphs of Local Invariants, *J. Mol. Struct.* (*Theochem*) **2001**, *538*, 287–193.

[18] A. A. Toropov and A. P. Toropova, Modeling of Lipophilicity by Means of Correlation Weighting of Local Graph Invariants, *J. Mol. Struct.* (*Theochem*) **2001**, *538*, 197–199.

[19] P. Peruzzo, D. M. Marino, A. A. Toropov and E. A. Castro, Calculation of pK Values of Flavilium Salts from the Optimization of Correlation Weights of Local Graph Invariants, *J. Mol. Struct.* (*Theochem*) **2001**, *572*, 53–60.

[20] A. Mercader, E. A. Castro and A. A. Toropov, Calculation of Total Molecular Electronics Energies from Correlations Weighting of Local Graph Invariants, *J. Mol. Model*. **2001**, *7*, 1–5.

[21] G. Krenkel, E. A. Castro and A. A. Toropov, Improved Molecular Descriptors to Calculate Boiling Points Based on the Optimization of Correlation Weights of Local Graph Invariants, *J. Mol. Struct.* (*Theochem*) **2001**, *542*, 107–113.

[22] G. Krenkel, E. A. Castro and A. A. Toropov, Improved Molecular Descriptors Based on the Optimization of Correlation Weights of Local Graph Invariants, *Int. J. Mol. Sci.***2001**, *2*, 57–65.

[23] A. Mercader, E. A. Castro and A. A. Toropov, QSPR Modeling of the Enthalpy of Formation from Elements by Means of Correlation Weighting of Local Invariants of Atomic Molecular Graphs, *Chem. Phys. Lett.* **2000**, *330*, 612–623.

[24] G. Krenkel, E. A. Castro and A. A. Toropov, 3D and 4D Molecular Models Derived from the Ideal Symmetry Method. Prediction of Alkanes Normal Boiling Points, *Chem. Phys. Lett.* **2002**, *355*, 517–528.

[25] D. J. G. Marino, P. J. Peruzzo, E. A. Castro, and A. A. Toropov, QSAR Carcinogenic Study of Methylated Polycyclic Aromatic Hydrocarbons Based on Topological Descriptors Derived from Distance Matrices and Correlation Weights of Local Graph Invariants, *Internet Electron. J. Mol. Des.* **2002**, *1*, 115–133, http://www.biochempress.com.

[26] P. Duchowicz, E. A. Castro and A. A. Toropov, Improved QSPR Analysis of Standard Entropy of Acyclic and Aromatic Compounds Using Optimized Correlation Weights of Linear Graph Invariants, *Comp. Chem.* **2002**, *26*, 327–332.

[27] S. C. Basak, G. J. Niemi, G. D. Veith, Predicting Properties of Molecules Using Graph Invariants, *J. Math. Chem.* **1991**, *7*, 243–272.

[28] V. R. Magnuson, D. K. Harriss and S. C. Basak, in *Chemical Applications of Topology and Graph Theory*, R. B. King, Ed., Elsevier, Amsterdam, 1983, p. 178.

[29] N. Trinajstic, *Chemical Graph Theory*, Vols. 1 and 2, CRC Press, Boca Raton, Fl, 1983.

[30] J. W. Kennedy and L. V. Quintas, *Applications of Graphs in Chemistry and Physics*, North–Holland, Amsterdam, 1988.

[31] D. Bonchev and O. Mekenyan, Eds., *Graph Theoretical Approaches to Chemical Reactivity*, Kluwer Academic Publishers, Dordrecht, 1994.

[32] A. T. Balaban, Ed., *Chemical Applications of Graph Theory*, Academic Press, New York, 1976.

[33] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures*, UMI, Bell & Howell Company, Ann Arbor, Michigan, 1983.

[34] M. Randic, *Topological Indices in Encyclopedia of Computational Chemistry*, P. v. R. Schleyer, Ed., Wiley & Sons, New York, 1998/1999, 3018–3032.

[35] M. Randic, On the Characterization of Molecular Branching, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

[36] M. Randic and S. C. Basak, Optimal Molecular Descriptors Based on Weighted Path Numbers, *J. Chem. Inf.*

Comput. Sci. **1999**, *39*, 261–266.

[37] A. Leo and D. Weininger, *CLOGP Version 3.2 User Reference Manual*, Medicinal Chemistry Project, Pomona College, Claremont, Ca., 1984.

[38] S. C. Basak and G. D. Grunwald, Estimation of Lipophilicity from Molecular Structural Similarity, *New J. Chem.* **1995**, *19*, 231–237.

[39] A. A. Toropov and A. P. Toropova, QSAR Modeling of Toxicity on Optimization of Correlation Weights of Morgan Extended Connectivity, *J. Mol. Struct.* (*Theochem*) **2002**, *578*, 129–134.

[40] A. A. Toropov and A. P. Toropova, QSAR Modeling of Mutagenicity Based on Graphs of Atomic Orbitals, *Internet Electron. J. Mol. Des.* **2002**, *1*, 108–114, http://www.biochempress.com.

[41] A. V. Pasyukov, M. I. Skvortsova, V. A. Palyulin and N. S. Zefirov, Optimal Selection of the Measure of Molecular Similarity in QSAR Studies, *Dokl. Chem.* **2000**, *374*, 224–227.

Biographies

Pablo J. Peruzzo is technical assistant at the Research Center of Environment Medium, La Plata University, and his main research interest is related to numerical methods to compute molecular descriptors in QSAR/QSPR field.

Damián J. G. Marino is technical assistant at the Research Center of Environment Médium, La Plata University, and is devoted to stuy carcinogenic hydrocarbons via theoretical and experimental methods.

Eduardo A. Castro is full professor at La Plata University and Superior Researcher belonging to the National Research Council. He has published around 700 papers and is author of several books and chapters of books on Physical Chemistry, Science Education, Mathematical Chemistry and Science Management issues.

Andrey A. Toropov is employed at the Vostok Innovation Company, Tashkent, Uzbekistan, and has published several papers on QSAR/QSPR theory. His main scientific interest is related to flexible topological molecular descriptors.