Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information

Lowell H. Hall*

Department of Chemistry, Eastern Nazarene College, Quincy, Massachusetts 02170

Lemont B. Kier

Department of Medicinal Chemistry, School of Pharmacy, Virginia Commonwealth University, Richmond, Virginia 23298

Received June 27, 1995[⊗]

The electrotopological state formalism is developed further in conjunction with atom classification. The classification scheme is based on the characteristics of hydride groups: (1) atomic number of an atom as element identifier, (2) a valence state designation consisting of valence and simple connectivity δ values (for each atom together with its bonded hydrogen atoms, as in $-CH_3$ or $-NH-$), and (3) an aromaticity indicator. This scheme may be viewed as a three-dimensional array. In a few cases, bonded neighbor analysis is also required. The scheme is developed and illustrated. For each atom type in a molecule, the electrotopological state indices are summed. These atom type E-state indices, based on a summation of E-state values, are useful for database characterization, molecular similarity analysis, and QSAR. A QSAR/QSPR example is given for boiling point for a set of 245 alkanes and alcohols for which the standard error is 8.0 °C.

INTRODUCTION

A new atom level topological index was introduced in 1990, called the electrotopological state index (E-state, for short). Other topological indices deal with the whole molecule as a sum over subgraphs of the molecular graph. In contrast, the E-state index is computed as a graph invariant for each atom in the molecular graph. The index combines the electronic state of the bonded atom within the molecule with its topological nature in the context of the whole molecular skeleton. The E-state indices have been used for a variety of QSAR studies.[1]

The atom type indices introduced here can be used in a manner similar to group additive schemes in which an index appears for each atom type in the molecule together with its contribution based on the E-state index. On the other hand, it may be possible that a limited number of atom type indices may be used for a given application, especially for biological data in which only a few atom types are required for a quality QSAR equation. For biological QSARs reported to date, a type of skeletal superposition is used so that the E-state values for corresponding atoms may be entered as variables in regression analysis.[1,3–8] The development of atom type E-state values provides the basis for application to a wider range of problems to which the E-state formalism is applicable without the need for superposition.

The paper describes the atom type classification scheme and the combination with the electrotopological state along with an illustration of the application of the resulting atom type E-state indices.

METHOD

This paper introduces an atom classification scheme in combination with the computed E-state index for the atoms in a molecule. In this manner, the sum of E-state indices can be accumulated for each atom type. The atom-type E-state sum indices, thus defined, can then be used in various applications in chemistry and drug research, including QSAR, database characterization, clustering, molecular similarity analysis, and related areas of investigation.

Electrotopological State Indices. A new paradigm, called the electrotopological state for atom electronic and topological characterization, was introduced by Kier and Hall and was reviewed recently.[1] For simplicity, these indices are referred to as the E-state. Each atom in the molecular graph is represented by an E-state variable which encodes the intrinsic electronic state of the atom as perturbed by the electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. Thus, the E-state for a given atom (type) varies from molecule to molecule and depends upon the detailed structure of the molecule. For example, the E-state value computed for an aromatic carbon atom is smaller when adjacent to a carbon substituted with an $-OH$ group and even smaller when the $-OH$ group is bonded directly to that carbon atom; both of these values are smaller than the E-state value for the unsubstituted carbon atom in benzene.

The intrinsic state is based on the Kier–Hall electronegativity[3] and derived from the ratio of that electronegativity to the number of skeletal sigma bonds for that atom[1]

$$I = ((2/N)^2\delta^v + 1)/\delta \qquad (1)$$

The symbol N is the principal quantum number for the valence shell of that atom; δ^v and δ are the molecular connectivity delta values which are given as follows (for first row atoms)

$$\delta = \sigma - h = \text{number of connections (edges) in the skeleton (graph)} \qquad (2)$$

[⊗] Abstract published in *Advance ACS Abstracts,* October 1, 1995.

Table 1. Atom Types, Atomic Numbers, Connectivity δ Values, Valence State Indicator, and Aromaticity Indicator Which Are Used as the Basis for a Classification Scheme for Atom Types in Conjunction with Electrotopological State Indices

		valence state indicator						
no.	atom group[a]	Z^b	$\delta^{v\,c}$	δ^d	$\delta^v + \delta^e$	$\delta^v - \delta^f$	AR[g]	group symbol[h]
1	−Li	3	1	1	2	0	0	sLi
2	−Be−	4	2	2	4	0	0	ssBe
3	>Be< [−2]	4	2	4	6	−2	0	ssssBe
4	−BH−	5	2	2	5	0	0	ssBH
5	>B−	5	3	3	6	0	0	sssB
6	>B< [−1]	5	3	4	7	−1	0	ssssB
7	−CH$_3$	6	1	1	2	0	0	sCH$_3$
8	=CH$_2$	6	2	1	3	1	0	dCH$_2$
9	−CH$_2$−	6	2	2	4	0[j]	0	ssCH$_2$
10	≡CH	6	3	1	4	2[j]	0	tCH
11	=CH−	6	3	2	5	1	0	dsCH
12	aCHa	6	3	2	5	1	1	aaCH
13	>CH−	6	3	3	6	0[j]	0	sssCH
14	=C=	6	4	2	6	2	0	ddC[i]
15	≡C−	6	4	2	6	2	0	tsC[i]
16	=C<	6	4	3	7	1	0	dssC
17	aCa−	6	4	3	7	1	1	aasC[i]
18	aaCa	6	4	3	7	1	1	aaaC[i]
19	>C<	6	4	4	8	0	0	ssssC
20	−NH$_3$ [+1]	7	2	1	3	1	0	sNH$_3$
21	−NH$_2$	7	3	1	4	2	0	sNH$_2$
22	−NH$_2$− [+1]	7	3	2	5	1	0	ssNH$_2$
23	=NH	7	4	1	5	3	0	dNH
24	−NH−	7	4	2	6	2	0	ssNH
25	aNHa	7	4	2	6	2	1	aaNH
26	≡N	7	5	1	6	4[j]	0	tN
27	>NH− [+1]	7	4	3	7	1	0	sssNH
28	=N−	7	5	2	7	3	0	dsN
29	aNa	7	5	2	7	3	1	aaN
30	>N−	7	5	3	8	2	0	sssN[i]
31	−N≪	7	5	3	8	2	0	ddsN (nitro)[i]
32	aaNs	7	5	3	8	2	0	aasN (N oxide)[i]
33	>N< [+1]	7	5	4	9	−1	0	ssssN (onium)
34	−OH	8	5	1	6	4	0	sOH
35	=O	8	6	1	7	5	0	dO
36	−O−	8	6	2	8	4	0	ssO
37	aOa	8	6	2	8	4	1	aaO
38	−F−	9	7	1	8	6	0	sF
39	−SiH$_3$	14	1	1	2	0	0	sSiH$_3$
40	−SiH$_2$−	14	2	2	4	0	0	ssSiH$_2$
41	>SiH−	14	3	3	6	0	0	sssSiH
42	>Si<	14	4	4	8	0	0	ssssSi
43	−PH$_2$	15	3	1	4	2	0	sPH$_2$
44	−PH−	15	4	2	6	2	0	ssPH
45	>P−	15	5	3	8	2	0	sssP
46	−>P=	15	5	4	9	1	0	dsssP
47	−>P<	15	5	5	10	0	0	sssssP
48	−SH	16	5	1	6	4	0	sSH
49	=S	16	6	1	7	5	0	dS
50	−S	16	6	2	8	4	0	ssS[i]
51	aSa	16	6	2	8	4	1	aaS
52	>S=	16	6	3	9	3	0	dssS (sulfone)
53	≥S≤	16	6	4	10	2	0	ddssS (sulfate)
54	−Cl	17	7	1	8	6	0	sCl
55	−GeH$_3$	32	1	1	2	0	0	sGeH$_3$
56	−GeH$_2$−	32	2	2	4	0	0	ssGeH$_2$
57	>GeH−	32	3	3	6	0	0	sssGeH
58	>Ge<	32	4	4	8	0	0	ssssGe
59	−AsH$_2$	33	3	1	4	2	0	sAsH$_2$
60	−AsH−	33	4	2	6	2	0	ssAsH
61	>As−	33	5	3	8	2	0	sssAs
62	−>As=	33	5	4	9	1	0	sssdAs
63	−>As<	33	5	5	10	0	0	sssssAs
64	−SeH	34	5	1	6	4	0	sSeH
65	=Se	34	6	1	7	5	0	dSe
66	−Se−	34	6	2	8	4	0	ssSe
67	aSea	34	6	2	8	4	1	aaSe
68	>Se=	34	6	3	9	3	0	dssSe
69	≥Se=	34	6	4	10	2	0	ddssSe
70	−Br	35	7	1	8	6	0	sBr
71	−SnH$_3$	50	1	1	2	0	0	sSnH$_3$
72	−SnH$_2$−	50	2	2	4	0	0	ssSnH$_2$
73	>SnH−	50	3	3	6	0	0	sssSnH
74	>Sn<	50	4	4	8	0	0	ssssSn
75	−I	53	7	1	8	6	0	sI
76	−PbH$_3$	82	1	1	2	0	0	sPbH$_3$
77	−PbH$_2$−	82	2	2	4	0	0	ssPbH$_2$
78	>PbH−	82	3	3	6	0	0	sssPbH
79	>Pb<	82	4	4	8	0	0	ssssPb

[a] Indication of atom groups according to valence state together with number of bonded hydrogens. Symbols: − for single, = for double, ≡ for triple, a for aromatic, ≪ for two double bonds or two resonance single/double bonds as in nitro group, ≥ for a single and a double bond, and −> for three single bonds. [b] Atomic number of element. [c] Valence connectivity δ value. [d] Simple connectivity δ value. [e] Sum of valence and simple δ value. [f] Difference of valence and simple δ value. [g] Indicator that atom is part of aromatic system: 1 indicates aromaticity. [h] Symbols for atom type as used in conjunction with the atom type electrotopological state index. [i] Denotes an atom which can be distinguished from another only by analysis of neighboring atoms. See text. [j] Cases in which the δ difference, $\delta^v - \delta$, is required as an additional classification criterion. See text.

where σ is the number of electrons in σ orbitals; h is the number of hydrogen atoms bonded to the atom

$$\delta^v = Z^v - h = \sigma + \pi + n - h \qquad (3)$$

where Z^v is the number of valence electrons, π is the number of electrons in π orbitals, and n is the number of electrons in lone pairs. It can be seen that the difference between the two δ values, $(\delta^v - \delta)$, is equal to the number of π and lone pair electrons which Kier and Hall showed is proportional to valence state electronegativity.[2] The intrinsic state value is large for electronegative atoms, especially for those with few skeletal connections, and is smaller for less electronegative atoms and for atoms with several σ bonds.

The E-state index for atom i, S_i, is defined as follows

$$S_i = I_i + \Delta I_i \qquad (4)$$

and the perturbation term

$$\Delta I_i = \sum (I_i - I_j)/r_{ij}^2 \quad \text{sum over all atoms } j \neq i \qquad (5)$$

where r_{ij} is the distance between atoms counted as the graph distance (d_{ij}) plus one. As a consequence of this definition, atoms which possess π and lone pair electrons or are terminal atoms or lie on the mantle of the molecule tend to have large positive values for S_i. Atoms which do not have π and lone pair electrons and/or are buried in the interior of the molecule tend to have small or negative E-state values. See Table 1 for a sample calculation as well as the references cited below.

The E-state indices have been used for ethers, aldehydes, and ketones to correlate ^{17}O NMR frequencies:[1,3,4] binding studies including binding of a series of indolealkylamines to 5-HT$_2$ receptors[1] and binding of barbiturates to β-cyclodextrin;[5] receptor binding QSAR including affinity of β-carbolines[6] and dopamine D-2 receptor binding of salicylamides;[7] and inhibition studies including inhibition of flu virus by benzimidazoles[5] and inhibition of MAO by hydrazides.[6] The MAO inhibition study was extensively developed to include a careful analysis of the inhibitor molecules using semiempirical MO computations. The model based on the E-state indices was found to be superior to that based on MO computed charges;[9] the time requirements for the MO study was at least 1000 times more than that required for the E-state analysis.

Atom Type Classification Scheme. In this classification scheme, each atom in the molecule is essentially identified by its valence state, including the number of attached hydrogen atoms. Such a grouping, an atom plus bonded hydrogen atoms, is sometimes called a hydride group. This classification is readily carried out by the Molconn software[9] and will be part of the forthcoming version 3.0.

Classification is based on four hydride group characteristics: (1) atom (element) identification; (2) valence state, including aromaticity indication; (3) number of bonded hydrogen atoms; and (4) in a few cases, the identity of other bonded atoms.

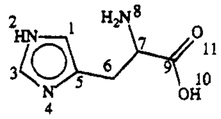
1. Atom (Element) Identity. This characteristic is directly represented by the atomic number, Z, which is an unambiguous element identifier.

2. Valence State. Valence state identification is somewhat less direct than element identification by atomic number. Valence state is usually understood to be based on the number of valence electrons assigned to σ, π, and lone pair orbitals. Also usually included is some representation of resonance forms and/or aromaticity. The hydrid state symbols such as sp^3, sp^2, sp, etc. are common and are widely used to describe valence state. However, we are interested in computer automated assignment of atom type; such hybrid state designations are not usually supplied as part of a connection table or other forms of molecule input data. We showed that the molecular connectivity δ values, δ^v and δ, are useful in the designation of valence state.[2] These δ values were defined in eqs 2 and 3. Further, their sum and difference values have also been shown to be related to van der Waals volume and to valence state electronegativity, respectively.[2] For encoding of a particular valence state, we have found that the δ value sum, $\delta^v + \delta$, is very useful as shown in Table 1. The δ value summation, $\delta^v + \delta$, will be called the valence state indicator (VSI). The completion of the valence state designation makes use of a marker for an atom which is part of an aromatic system as shown in Table 1. For our purposes we use a simple bivariate indicator, AR: 1 for aromatic, 0 for nonaromatic. This indicator may be directly supplied, as in the case of lower case letters in SMILES strings, but may need to be determined from the Hückel rule otherwise.

3. Number of Bonded Hydrogen Atoms. It has proven useful in various group additive schemes to create separate classes for atoms in the same valence state but with a different number of hydrogen atoms, for example, $-CH_3$, $-CH_2-$, $-CH<$, and $>C<$; $-NH_2$, $-NH-$, and $-N<$; $-OH$, and $-O-$, etc. Hydrogen atoms are important because they may be involved in intermolecular contacts and interactions, and, further, they contribute to overall property values. For this reason we consider each of these hydride groups as different classes. Examination of Table 1 shows that the δ sum variable, $\delta^v + \delta$, classifies such groups approximately in addition to valence state designation (with a few exceptions as described below).

4. Distinguishing Identity of Bonded Atoms. There are a few cases for which the atomic number and the valence state designation do not distinguish among atoms which clearly belong to different groups. For the groups considered in this paper, the following groups require bonded-atom analysis: allenic and acetylenic carbon atoms; a carbon atom at the juncture of two aromatic rings as in the 9 and 10 positions in naphthalene and an aromatic carbon which is bonded to a substituent such as the *ipso* atom in 1-naphthol; the tertiary amine nitrogen, the nitro nitrogen, and the nitrogen in a pyridine *N*-oxide; and the sulfur in a sulfide and the sulfur in a disulfide link. In these four cases, use is made of the adjacency matrix to determine the nature of bonded atoms as a basis for assignment to the appropriate group. For example, the 9 and 10 carbons in naphthalene are bonded to three atoms with aromatic markers whereas the aromatic carbon with an external single bond is bonded to only two atoms with aromatic markers.

Table 2. An Illustration of the Computed E-State Values for the Atoms Present in the Molecule Along with the Summation Atom Type E-State Indices for the Atom Types Present in the Molecule



atom no.	atom type symbol	electrotopological state index value
1	aaCH	1.628
2	aaNH	2.714
3	aaCH	1.490
4	aaN	3.840
5	aasC	0.666
6	$ssCH_2$	0.263
7	sssCH	−0.863
8	sNH_2	5.257
9	dssC	−1.006
10	sOH	8.418
11	dO	10.263

atom type E-state symbol	sum index value
SdO	10.263
SsOH	8.418
$SsNH_2$	5.257
SaaN	3.840
SaaNH	2.714
SaaCH	2.118
SaasC	0.666
SsssCH	−0.863
SdssC	−1.006

This classification scheme may be viewed as a three-dimensional array with the atomic number as one dimension, the δ sum ($\delta^v + \delta$) as the second, and the aromatic indicator as the third dimension. This array correctly classifies all but 10 of the 79 groups listed in Table 1. Six are properly classified by the bonded-atom analysis just described. They are marked with a single asterisk in Table 1. The remaining four cases are resolved by use of the δ difference, $\delta^v - \delta$, as shown in Table 2 where they are marked with a double asterisk. For example, the triply bonded nitrogen atom, $\equiv N$, has $\delta^v - \delta = 4$, whereas $=N-$ has $\delta^v - \delta = 3$. In each of these cases, the δ sum is the same for two groups but the δ difference is different.

In this manner, a modified version of Molconn-X (version 2.0) classifies each atom in the molecule.[9] The count of each atom type is accumulated along with the sum of E-state values for all groups of the same type. These two sets of

Table 3. Boiling Point of Alkyl Alcohols (with 2–10 Carbon Atoms) and Alkanes (with 5–10 Carbon Atoms) Modeled with the Atom-Type Electrotopological State Indices

		boiling point (°C)			
id	name	obs[a]	calc[b]	res[c]	pres[d]
1	ethanol	78.0	61.6	16.4	17.5
2	propanol	97.1	81.3	15.8	16.6
3	isopropyl alcohol	82.4	74.0	8.4	8.8
4	butanol	117.6	102.2	15.4	16.0
5	2-methyl-1-propanol	108.1	98.6	9.5	9.9
6	2-butanol	99.5	93.9	5.6	5.8
7	2-methyl-2-propanol	82.4	76.3	6.1	6.4
8	pentanol	138.0	123.6	14.4	14.9
9	3-methyl-1-butanol	131.0	121.0	10.0	10.3
10	2-methyl-1-butanol	128.0	119.7	8.3	8.5
11	2-pentanol	119.3	114.7	4.6	4.6
12	3-pentanol	116.2	114.2	2.0	2.1
13	3-methyl-2-butanol	112.9	110.3	2.6	2.7
14	2-methyl-2-butanol	102.3	96.6	5.7	5.9
15	hexanol	157.6	145.3	12.3	12.7
16	3-methyl-1-pentanol	153.0	142.6	10.4	10.6
17	4-methyl-1-pentanol	151.9	143.0	8.9	9.1
18	2-methyl-1-pentanol	149.0	141.1	7.9	8.1
19	2-ethyl-1-butanol	147.0	141.2	5.8	5.9
20	2,,3-dimethyl-1-butanol	144.5	137.6	6.9	7.1
21	3,3-dimethyl-1-butanol	143.0	137.0	6.0	6.2
22	2-hexanol	140.0	136.1	3.9	4.0
23	2,2-dimethyl-1-butanol	136.5	132.8	3.7	3.8
24	3-hexanol	135.0	135.2	−0.2	−0.2
25	3-methyl-2-pentanol	134.3	131.4	2.9	2.9
26	4-methyl-2-pentanol	131.6	133.0	−1.4	−1.4
27	2-methyl-3-pentanol	126.5	130.5	−4.0	−4.1
28	3-methyl-3-pentanol	122.4	117.3	5.1	5.3
29	2-methyl-2-pentanol	121.1	117.4	3.7	3.8
30	3,3-dimethyl-2-butanol	120.4	121.4	−1.0	−1.0
31	2,3-dimethyl-2-butanol	118.4	111.4	7.0	7.3
32	heptanol	176.4	167.1	9.3	9.5
33	4-methyl-1-hexanol	173.0	165.0	8.0	8.2
34	5-methyl-1-hexanol	170.0	165.0	5.0	5.1
35	3-methyl-1-hexanol	169.0	164.3	4.7	4.8
36	2-methyl-1-hexanol	164.0	162.7	1.3	1.3
37	2-heptanol	159.0	157.7	1.3	1.3
38	2,4-dimethyl-1-pentanol	159.0	159.9	−0.9	−1.0
39	3-heptanol	157.0	156.6	0.4	0.4
40	4-heptanol	156.0	156.2	−0.2	−0.2
41	5-methyl-2-hexanol	151.0	155.1	−4.1	−4.2
42	5-methyl-3-hexanol	148.0	153.3	−5.3	−5.4
43	2-methyl-2-hexanol	143.0	138.6	4.4	4.5
44	2-methyl-3-hexanol	143.0	151.5	−8.5	−8.6
45	3-methyl-3-hexanol	143.0	138.2	4.8	4.9
46	3-ethyl-3-pentanol	142.0	138.3	3.7	3.8
47	2,3-dimethyl-3-pentanol	139.7	132.1	7.6	7.9
48	2,4-dimethyl-3-pentanol	138.7	146.3	−7.6	−7.8
49	2,2-dimethyl-3-pentanol	135.0	141.5	−6.5	−6.7
50	2,4-dimethyl-2-pentanol	133.1	134.6	−1.5	−1.6
51	2,3,3-trimethyl-2-butanol	131.0	119.8	11.2	12.2
52	octanol	195.1	189.1	6.0	6.2
53	6-methyl-1-heptanol	188.6	187.0	1.6	1.6
54	4-methyl-1-heptanol	188.0	186.8	1.2	1.2
55	2-octanol	180.0	179.5	0.5	0.5
56	2,5-dimethyl-1-hexanol	179.5	182.1	−2.6	−2.6
57	4-octanol	176.3	177.7	−1.4	−1.4
58	6-methyl-3-heptanol	174.0	175.5	−1.5	−1.6
59	5-methyl-3-heptanol	172.0	175.0	−3.0	−3.0
60	3-octanol	171.0	178.3	−7.3	−7.4
61	5-methyl-2-heptanol	170.0	177.0	−7.0	−7.1
62	4-methyl-3-heptanol	170.0	173.2	−3.2	−3.3
63	2,4,4-trimethyl-1-pentanol	168.5	176.3	−7.8	−8.0
64	2-methyl-3-heptanol	167.5	172.8	−5.3	−5.4
65	3-methyl-2-heptanol	166.1	174.3	−8.2	−8.3
66	3,4-dimethyl-2-hexanol	165.5	170.4	−4.9	−5.0
67	2-methyl-4-heptanol	164.0	174.3	−10.3	−10.5
68	3-methyl-3-heptanol	163.0	159.5	3.5	3.6
69	3-methyl-4-heptanol	162.0	172.8	−10.8	−11.0
70	4-methyl-4-heptanol	161.0	159.2	1.8	1.8
71	2-methyl-3-ethyl-3-pentanol	160.0	153.0	7.0	7.2
72	2,3-dimethyl-2-hexanol	160.0	153.7	6.3	6.6
73	2,3,4-trimethyl-3-pentanol	156.5	146.3	10.2	10.8
74	2-methyl-3-ethyl-2-pentanol	156.0	153.8	2.2	2.2
75	2-methyl-2-heptanol	156.0	160.1	−4.1	−4.3
76	2,5-dimethyl-2-hexanol	154.5	157.0	−2.5	−2.5
77	2,2,4-trimethyl-3-pentanol	150.5	156.3	−5.8	−6.0
78	2,4,4-trimethyl-2-pentanol	147.5	148.0	−0.5	−0.5
79	nonanol	213.3	211.1	2.2	2.3
80	7-methyloctanol	206.0	209.1	−3.1	−3.2
81	3-nonanol	195.0	200.1	−5.1	−5.2
82	2-nonanol	193.5	201.5	−8.0	−8.2
83	5-nonanol	193.0	199.2	−6.2	−6.4
84	4-nonanol	192.5	199.4	−6.9	−7.1
85	4-ethyl-4-heptanol	182.0	180.5	1.5	1.5
86	2-methyl-2-octanol	178.0	181.8	−3.8	−4.0
87	2,6-dimethyl-3-heptanol	175.0	191.5	−16.5	−16.9
88	2,6-dimethyl-4-heptanol	174.5	192.1	−17.6	−18.1
89	2,6-dimethyl-2-heptanol	173.0	179.1	−6.1	−6.3
90	3,6-dimethyl-3-heptanol	173.0	177.8	−4.8	−5.0
91	2,2,3-trimethyl-3-hexanol	156.0	161.1	−5.1	−5.6
92	decanol	231.1	233.2	−2.1	−2.3
93	3,7-dimethyl-1-octanol	212.5	227.7	−15.2	−15.9
94	2-decanol	211.0	223.5	−12.5	−13.0
95	4-decanol	210.5	221.3	−10.8	−11.2
96	3,6-dimethyl-3-octanol	202.2	199.6	2.6	2.7
97	3-ethyl-3-octanol	199.0	202.3	−3.3	−3.5
98	2,6-dimethyl-4-octanol	195.0	213.8	−18.8	−19.4
99	2,7-dimethyl-3-octanol	193.5	213.6	−20.1	−20.7
100	3-ethyl-2-methyl-3-heptanol	193.0	195.4	−2.4	−2.5
101	pentane	36.1	53.0	−16.9	−17.9
102	2-methylbutane	27.9	51.8	−23.9	−24.9
103	2,2-dimethylpropane	9.5	49.8	−40.3	−42.3
104	hexane	68.7	75.0	−6.3	−6.6
105	2-methylpentane	60.3	73.6	−13.3	−13.7
106	3-methylpentane	63.3	73.9	−10.6	−10.9
107	2,3-dimethylbutane	58.0	71.5	−13.5	−14.0
108	2,2-dimethylbutane	49.7	72.1	−22.4	−23.2
109	heptane	98.4	97.1	1.3	1.3
110	2-methylhexane	90.1	95.6	−5.5	−5.6
111	3-methylhexane	91.8	95.9	−4.1	−4.2
112	3-ethylpentane	93.5	96.4	−2.9	−3.0
113	2,2-dimethylpentane	79.2	93.8	−14.6	−15.0
114	2,3-dimethylpentane	89.8	93.7	−3.9	−4.0
115	2,4-dimethylpentane	80.5	93.7	−13.2	−13.5
116	3,3-dimethylpentane	86.1	94.6	−8.5	−8.7
117	2,2,3-trimethylbutane	80.9	90.3	−9.4	−9.6
118	octane	125.7	119.3	6.4	6.7
119	2-methylheptane	117.6	117.6	0.0	−0.0
120	3-methylheptane	118.9	118.0	0.9	0.9
121	4-methylheptane	117.7	118.0	−0.3	−0.3
122	3-ethylhexane	118.5	118.5	−0.0	−0.1
123	2,2-dimethylhexane	106.8	115.7	−8.9	−9.1
124	2,3-dimethylhexane	115.6	115.6	0.0	−0.0
125	2,4-dimethylhexane	109.4	115.9	−6.5	−6.6
126	2,5-dimethylhexane	109.1	115.8	−6.7	−6.8
127	3,3-dimethylhexane	112.0	116.6	−4.6	−4.7
128	3,4-dimethylhexane	117.7	116.0	1.7	1.8
129	2,2,3-trimethylpentane	109.8	112.4	−2.6	−2.6
130	2,2,4-trimethylpentane	99.2	112.9	−13.7	−13.9
131	2,3,3-trimethylpentane	114.8	112.8	2.0	2.0
132	2,3,4-trimethylpentane	113.5	112.9	0.6	0.7
133	2-methyl-3-ethylpentane	115.7	116.1	−0.4	−0.4
134	3-ethyl-3-methylpentane	118.3	117.5	0.8	0.8
135	2,2,3,3-tetramethylbutane	106.5	106.3	0.2	0.2
136	nonane	150.8	141.5	9.3	9.8
137	2-methyloctane	143.3	139.7	3.6	3.7
138	3-methyloctane	144.2	140.1	4.1	4.2
139	4-methyloctane	142.5	140.1	2.4	2.4
140	3-ethylheptane	143.0	140.7	2.3	2.3
141	4-ethylheptane	141.2	140.8	0.4	0.4
142	2,2-dimethylheptane	132.7	137.6	−4.9	−5.0
143	2,3-dimethylheptane	140.5	137.6	2.9	2.9
144	2,4-dimethylheptane	133.5	137.9	−4.4	−4.5

Table 3 (Continued)

		boiling point (°C)						boiling point (°C)			
id	name	obs[a]	calc[b]	res[c]	pres[d]	id	name	obs[a]	calc[b]	res[c]	pres[d]
145	2,5-dimethylheptane	136.0	138.1	−2.1	−2.1	196	2-methyl-3-ethylheptane	166.0	160.3	5.7	5.8
146	2,6-dimethylheptane	135.2	137.9	−2.7	−2.7	197	2-methyl-4-ethylheptane	160.0	160.6	−0.6	−0.7
147	3,3-dimethylheptane	137.3	138.5	−1.2	−1.2	198	2-methyl-5-ethylheptane	159.7	160.7	−1.0	−1.0
148	3,4-dimethylheptane	140.6	138.0	2.6	2.6	199	3-methyl-4-ethylheptane	167.0	160.8	6.2	6.3
149	3,5-dimethylheptane	136.0	138.2	−2.2	−2.3	200	3-methyl-5-ethylheptane	161.0	160.9	0.1	0.1
150	4,4-dimethylheptane	135.2	138.6	−3.4	−3.5	201	4-methyl-3-ethylheptane	167.0	160.7	6.3	6.4
151	2,2,3-trimethylhexane	133.6	134.1	−0.5	−0.5	202	4-propylheptane	162.0	163.1	−1.1	−1.2
152	2,2,4-trimethylhexane	126.5	135.0	−8.5	−8.6	203	2,2,3-trimethylheptane	158.0	156.0	2.0	2.0
153	2,2,5-trimethylhexane	124.1	135.2	−11.1	−11.2	204	2,2,4-trimethylheptane	147.7	156.9	−9.2	−9.3
154	2,3,3-trimethylhexane	137.7	134.7	3.0	3.1	205	2,2,5-trimethylheptane	148.0	157.4	−9.4	−9.5
155	2,3,4-trimethylhexane	139.0	135.1	3.9	4.0	206	2,2,6-trimethylheptane	148.2	157.4	−9.2	−9.3
156	2,3,5-trimethylhexane	131.3	135.2	−3.9	−4.0	207	2,3,3-trimethylheptane	160.0	156.6	3.4	3.5
157	3,3,4-trimethylhexane	143.5	135.0	8.5	8.6	208	2,4,4-trimethylheptane	153.0	157.5	−4.5	−4.6
158	2,4,4-trimethylhexane	130.6	135.6	−5.0	−5.0	209	2,5,5-trimethylheptane	152.8	158.0	−5.2	−5.2
159	2-methyl-3-ethylhexane	138.0	138.2	−0.2	−0.2	210	3,3,4-trimethylheptane	164.0	157.0	7.0	7.1
160	2-methyl-4-ethylhexane	133.8	138.5	−4.7	−4.7	211	3,3,5-trimethylheptane	155.7	157.8	−2.1	−2.1
161	3-methyl-3-ethylhexane	140.6	139.7	0.9	1.0	212	3,4,4-trimethylheptane	164.0	157.0	7.0	7.1
162	3-methyl-4-ethylhexane	140.4	138.6	1.8	1.8	213	3-methyl-3-ethylheptane	163.8	161.6	2.2	2.2
163	2,2-dimethyl-3-ethylpentane	133.8	134.7	−0.9	−0.9	214	4-methyl-4-ethylheptane	167.0	161.8	5.2	5.3
164	2,3-dimethyl-3-ethylpentane	142.0	135.7	6.3	6.4	215	2,2-dimethyl-3-ethylhexane	159.0	156.7	2.3	2.3
165	2,4-dimethyl-3-ethylpentane	136.7	135.3	1.4	1.4	216	2,2-dimethyl-4-ethylhexane	147.0	157.4	−10.4	−10.5
166	2,2,3,3-tetramethylpentane	140.3	128.7	11.6	12.0	217	2,3-dimethyl-4-ethylhexane	164.0	157.6	6.4	6.6
167	2,2,3,4-tetramethylpentane	133.0	130.5	2.5	2.6	218	2,4-dimethyl-3-ethylhexane	164.0	157.7	6.3	6.5
168	2,2,4,4-tetramethylpentane	122.3	130.5	−8.2	−8.6	219	2,5-dimethyl-3-ethylhexane	157.0	157.7	−0.7	−0.8
169	2,3,3,4-tetramethylpentane	141.6	130.4	11.2	11.4	220	3,4-diethylhexane	162.0	161.3	0.7	0.7
170	3,3-diethylpentane	146.2	140.8	5.4	5.5	221	4-isopropylheptane	160.0	160.4	−0.4	0.4
171	decane	174.1	163.7	10.4	11.1	222	2-methyl-3-isopropylhexane	163.0	157.3	5.7	5.8
172	2-methylnonane	166.8	161.9	4.9	5.1	223	2,2,3,3-tetramethylhexane	160.3	150.5	9.8	10.2
173	3-methylnonane	167.8	162.2	5.6	5.8	224	2,2,3,4-tetramethylhexane	154.9	152.6	2.3	2.3
174	4-methylnonane	165.7	162.2	3.5	3.5	225	2,2,3,5-tetramethylhexane	148.4	153.0	−4.6	−4.7
175	5-methylnonane	165.1	162.2	2.9	2.9	226	2,2,4,4-tetramethylhexane	153.3	153.1	0.2	0.2
176	2,2-dimethyloctane	155.0	159.6	−4.6	−4.7	227	2,2,4,5-tetramethylhexane	147.9	153.6	−5.7	−5.9
177	2,3-dimethyloctane	163.8	159.6	4.2	4.2	228	2,2,5,5-tetramethylhexane	137.5	153.5	−16.0	−16.7
178	2,4-dimethyloctane	153.0	160.0	−7.0	−7.1	229	2,3,3,4-tetramethylhexane	164.6	152.6	12.0	12.3
179	2,5-dimethyloctane	158.0	160.2	−2.2	−2.2	230	2,3,3,5-tetramethylhexane	153.0	153.3	−0.3	−0.3
180	2,6-dimethyloctane	158.5	160.2	−1.7	−1.8	231	2,3,4,4-tetramethylhexane	162.2	153.1	9.1	9.2
181	2,7-dimethyloctane	159.9	160.0	−0.1	−0.1	232	2,3,4,5-tetramethylhexane	161.0	153.8	7.2	7.6
182	3,3-dimethyloctane	161.2	160.5	0.7	0.8	233	3,3,4,4-tetramethylhexane	170.0	151.4	18.6	19.4
183	3,4-dimethyloctane	166.0	160.0	6.0	6.1	234	2,4-dimethyl-4-ethylhexane	158.0	158.5	−0.5	−0.5
184	3,5-dimethyloctane	160.0	160.3	−0.3	−0.3	235	3,3-dimethyl-4-ethylhexane	165.0	157.6	7.4	7.5
185	3,6-dimethyloctane	160.0	160.5	−0.5	−0.5	236	2,3-dimethyl-3-ethylhexane	169.0	157.7	11.3	11.4
186	4,4-dimethyloctane	161.0	160.6	0.4	0.4	237	3,4-dimethyl-3-ethylhexane	170.0	158.1	11.9	12.1
187	4,5-dimethyloctane	162.1	160.1	2.0	2.1	238	2,2,3-trimethyl-3-ethylpentane	168.0	151.6	16.4	17.1
188	3-ethyloctane	168.0	162.8	5.2	5.3	239	3,3-dimethylhexane	166.3	163.0	3.3	3.4
189	4-ethyloctane	168.0	163.0	5.0	5.2	240	2-methyl-3,3-diethylpentane	174.0	158.9	15.1	15.3
190	2,3,4-trimethylheptane	163.0	157.0	6.0	6.1	241	2,2,4-trimethyl-3-ethylpentane	155.3	152.9	2.4	2.5
191	2,3,5-trimethylheptane	157.0	157.5	−0.5	−0.5	242	3-ethyl-2,3,4-trimethylpentane	169.4	153.3	16.1	16.5
192	2,3,6-trimethylheptane	155.7	157.4	−1.7	−1.8	243	2,4-dimethyl-3-isopropylpentane	157.0	153.9	3.1	3.3
193	2,4,5-trimethylheptane	157.0	157.5	−0.5	−0.5	244	2,2,3,3,4-pentamethylpentane	166.1	145.3	20.8	21.7
194	2,4,6-trimethylheptane	144.8	157.6	−12.8	−13.1	245	2,2,3,4,4-pentamethylpentane	159.3	146.5	12.8	13.4
195	3,4,5-trimethylheptane	164.0	157.4	6.6	6.8						

[a] Normal boiling point, taken from CRC Handbook of Chemistry and Physics, 75th ed.; David R. Lide, Ed.; CRC Press, Boca Raton, FL, 1994. [b] Computed with eq 6. [c] Observed − calculated. [d] Predicted residual, computed from $n - 1$ observations when the current observation is left out; used as a basis for the press statistics.

indices are stored and made available for the user. In principle, this scheme can be extended to other elements in their various hydride groups. Those groups shown in Table 1 are the ones currently implemented in version 3.0 of Molconn-X.

Table 1 gives a list of the atom types used in this paper along with the bond symbols used. For example, in the symbol SssCH2, 'S' stands for the sum of E-state values for all the $-CH_2-$ groups in the molecule, 'ss' stands for the two single bonds of that group, and 'CH_2' represents the formula of the hydride group. In this manner, it is possible to distinguish between $-CH_2-$ and $=CH_2$. Also, SaaCH stands for sum of E-state indices for the CH in an aromatic ring, and SsOH stands for the sum of E-state indices for $-OH$ groups in the molecule. Table 2 gives the E-state atom values as well as sum atom type E-state values for an example.

OSPR Example. To illustrate the applicability of the atom type E-State indices, we created a combined data set of alcohols and alkanes along with their boiling points, as shown in Table 3. We selected alcohols with two to ten carbon atoms[10] together with all the alkane isomers with five to ten carbon atoms.[11] The data set contains 245 compounds. The boiling points range from 9.5 °C (2,2-dimethylpropane) up to 231.0 °C (decanol). Multiple linear regression was run using the SAS system.[12] Five variables were entered into the data set: $SsCH_3$, $SssCH_2$, SsssCH, SssssC, and SsOH, representing all five of the atom types present in the data set, as defined above.

The regression equation is as follows

$$bp = 8.21\ (\pm 0.30)\ SsCH_3 + 14.86\ (\pm 0.28)\ SssCH_2 + 24.56\ (\pm 0.99)\ SsssCH + 43.76\ (\pm 2.58)\ SssssC + 11.63\ (\pm 0.22)\ SsOH - 43.95\ (\pm 3.53) \quad (6)$$

$$r = 0.97,\quad s = 8.0,\quad F = 755,\quad n = 245$$

The number in parentheses is the standard deviation associated with each coefficient. Each coefficient is highly significant; the t values for the coefficients range from 16.9 to 53.8.

For cross-validation, we use the PRESS statistics, obtained by leaving out each observation and predicting its value from the remaining $n - 1$ observation

$$r_{press} = 0.97,\quad s_{press} = 8.3\ °C$$

The mean absolute error, mae, is 5.9 °C which corresponds to a 4.1% relative error.

The observed, predicted, residual, and press values are given in Table 3. Examination of the residuals (as a function of observed bp) shows a generally random scatter.

DISCUSSION

The regression model developed for the boiling points of alcohols and alkanes, eq 6, is excellent. The standard deviation, 8.0 °C, is better than those usually obtained for data sets of this size and inclusiveness of isomers. It is this statistical quantity which is the most important because it relates directly to the interests of the experimental scientist who wishes to make a prediction. This is also the first report for this particular data set which includes available alcohols and all alkane isomers, as described. For comparison, regression with the single variable, total number of atoms per molecule yield s = 20.3 °C; for the three variables, number of carbon, hydrogen, and oxygen atom s = 11.6 °C; for the five variables, simple counts of each atom type present ($CsCH_3$, $CssCH_2$, CsssCH, CssssC, CsOH) s = 10.2 °C (using the same atom types as used in eq 6 but merely their counts). The 8.0 standard error found with the atom type E-state model is significantly better.

Further, examination of the plot of observed boiling point versus calculated boiling point for the model based on simple counts shows very clear separation of the two subsets of structures, alcohols and alkanes. This effect is clearly a major part of the 10.3 °C standard error. However, for the model based on atom type E-states, no such separation is found. There is also strong indication of nonlinearity found in the plot of residuals versus boiling point for the simple counts model but only slight indication for the atom type E-state model.

The quality of the model for use in predictions is indicated by the press statistic, s_{press} = 8.3 °C, only 0.3 °C higher than for the direct model. Such a small increase, 3.8%, indicates a strong model.

It is also possible to discuss the contribution to boiling point for each atom type. That contribution, for each compound, is given by the product of the regression coefficient and the numerical value of the variable. For example, for methyl groups in a given compound, the contribution is $8.2163*SsCH_3$. The contribution per methyl group is obtained from division by the count of methyl groups in the molecule, $CcCH_3$: $8.2163*SsCH_3/CsCH_3$. An analysis for this data set shows the following average contributions and ranges: for $-CH_3$, 21.89 with a range from 13.81 to 37.30; for $-CH_2-$, 17.82, with a range from 3.72 to 31.29; for $>CH-$, 16.36, with a range from −4.94 to 25.25; for $>C<$, 11.09, with a range from −21.88 to 29.54; and for −OH, 103.95, with a range from 88.06 to 116.67 C. The negative values occur in compounds in which the −OH group is bonded to a $>CH-$ or a $>C<$ group.

Atoms in a molecule are clearly characterized by the atom type E-state indices. A methyl group is distinguished from a methylene group. Further, the SsOH index varies according to the electronic and topological environment of the −OH group. It is also zero for the alkanes. In this fashion presence and absence of the −OH group is indicated and its varying character encoded when present. Therefore, molecules with different functional groups can be distinguished.

The fact that the contribution for each atom type is not a constant but, in fact, covers a significant range of values clearly shows why the simple group additive scheme gives a higher standard error for the regression. The E-state formalism gives to each group a value which is dependent upon the structural (bonding) environment of that group. The −OH group makes a greater contribution to boiling point when it is a secondary alcohol and greater still when tertiary. However, it must be pointed out that in the tertiary alcohol, the quaternary carbon (to which the −OH is attached) makes a negative contribution. The net result is that tertiary alcohols have a lower boiling point than other alcohols with similar skeletal structure. Likewise, methyl groups make larger contributions when located on a quaternary carbon or a methine; when located on a methylene group, the methyl group makes its smallest typical contribution although the difference is small, 18.6 compared to 18.9.

The nonlinear aspect of the relationship between molecular structure and boiling point has not been discussed here because we are actively developing a more general approach to nonlinearity than by the use of explicit nonlinear terms in multiple linear regression. It is this nonlinearity that is responsible for the few large residuals reported in Table 3. Future papers will report the use of the atom type E-state indices in the more general nonlinear modeling in which these large residuals do not occur.

CONCLUSIONS

The electrotopological state indices have demonstrated considerable usefulness in the establishment of QSAR/QSPR equations. The ability to focus on a small number of atoms has provided significant utility in their applicability. The fact that they encode important electronic and topological information invests in them the ability to portrary significant pharmacological information for database characterization.

The addition of atom typing extends the usefulness of the E-state indices so that broader data sets may be examined. The atom-by-atom match-up (or superposition) required when individual E-state values are used, which is so useful for biological studies, is not sensible for general physicochemical properties, such as boiling point. Atom groups do not present themselves to each other in the neat or solution state in a strict atom-by-atom matching arrangement. Rather, molecules present atom features to each other in the liquid state in a random fashion. Thus, the atom groups on the molecular surface make significantly greater contributions along with those features which are polar or participate in hydrogen bonding.

For the set of alcohols and alkanes, the atom type E-state indices provide a good multiple linear regression model. The intermolecular effects mentioned above are clearly seen in this study for the average contribution by group: $[-OH] \gg [-CH_3] > [-CH_2-] > [-CH<] > [>C<]$. Of course, this is not a new finding nor is it surprising; the atom type E-state values simply reflect this reality in the data set. However, the atom type E-state formalism does not require that the contribution of a group be constant; rather, the contribution depends upon the environment of the group within the particular molecule. This formalism gives a more accurate and chemically meaningful expression to the role of groups in molecules.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. An Atom-Centered Index for Drug QSAR Models. In *Advances in Drug Design*; Testa, B., Ed.; Academic Press: 1992; Vol. 22.

(2) Kier, L. B.; Hall, L. H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, *70*, 583−589.

(3) Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7*, 801−807.

(4) Hall, L. H.; Kier, L. B. An Index of Electrotopological State for Atoms in Molecules. *J. Math. Chem.* **1991**, *7*, 229−241.

(5) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(6) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat.* **1991**, *10*, 43−51.

(7) Hall, L. H.; Kier, L. B. Binding of Salicylamides: QSAR Analysis with Electrotopological State Indexes. *Med. Res. Rev.* **1992**, *2*, 497−502.

(8) Hall, L. H.; Mohney, B. K.; Kier, L. B. Comparison of Electrotopological State Indexes with Molecular Orbital Parameters: Inhibition of MAO by Hydrazides. *Quant Struct.-Act. Relat.* **1993**, *12*, 44−48.

(9) The program Molconn-X was used for computation of electrotopological state indices. Contact author L. H. Hall for information.

(10) CRC Handbook of Chemistry and Physics, 75th ed.; Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 1994.

(11) Physical Properties of Compounds-II, Advances in Chemistry Series, No. 22, Dreisbach, R. R., Ed.; American Chemical Society: Washington, DC, 1959.

(12) SAS Institute, Cary, NC.

CI950067J