# The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs

LOWELL H. HALL* and BRIAN MOHNEY

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

LEMONT B. KIER

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

Received April 5, 1990

The electrotopological state, a novel representation of atoms in molecules, is developed from chemical graph theory as an index of the graph vertex (or skeletal group). This new index combines both the electronic character and the topological environment of each skeletal atom in a molecule. The electrotopological state (E-state) of a skeletal atom is formulated as an intrinsic value $I_i$ plus a perturbation term, $\Delta I_i$, arising from the electronic interaction and modified by the molecular topological environment of each atom in the molecule. The atom intrinsic value, for first-row atoms, is given as $I = (\delta^v + 1)/\delta$, in which $\delta^v$ and $\delta$ are the counts of valence and $\sigma$ electrons, respectively, for the atom in the molecular skeleton. The E-state, $S_i$, for atom $i$ is defined as $S_i = I_i + \Delta I_i$, where the influence of other atoms on atom $i$, $\Delta I_i$, is given as $\sum(I_i - I_j)/r_{ij}^2$; $r_{ij}$ is the graph separation between atoms $i$ and $j$, counted as the number of atoms inclusive of $i$ and $j$. Information in the electrotopological state is revealed by examples of various types of organic structures, including skeletal branching and heteroatom variation. Applications of this new method are given for $^{17}$O NMR chemical shift and inhibition of flu virus.

## INTRODUCTION AND BACKGROUND

A central concept in chemistry is that properties of a molecule are intimately related to its molecular structure. The structure of the molecule includes the number and kinds of atoms and the connections between them.[1] The molecular skeleton is often used as a convenient vehicle for this structure information. It has been shown in recent years that much valuable information in structure-property relations can be obtained from the set of connections in the structure.[2-11] Although three-dimensional topography is important for certain aspects of molecular properties, the basic information resident in the molecular skeleton is sufficient for many useful relationships. Further, it is the connections within the skeleton, the bonding scheme, which determines the three-dimensional architecture.

In our work we have been exploring the structure information associated with the molecular skeleton. This information is often called topological. One approach to modeling chemical and biological properties is the encoding of molecular structure information in structure indexes. Several types of indexes have been developed, and all have been termed topological indexes. However, these indexes do not all encode the same attributes of structure, and the encoding is not done in the same manner. Whatever specific terms may be used to describe these encoded structure attributes, most of them resolve into what is generally called electronic or topological factors. Electronic factors include polarity, charge, energy levels, and others. Topological factors include the arrangement of atoms across the skeleton (skeletal connectedness and branching), concepts of steric relations and bulk, and the relationships between various nonbonded parts of the molecule.

Both of these categories of structure attributes are intrinsic to the structure of a molecule and one approach to structure index development seeks to embrace both in a unified model. It is anticipated that this union of molecular electronic and topological information into a single quantitative description is capable of modeling measured properties or activities. Such an approach is described as a unified attribution model. This model unifies structure information and is not based on measured molecular properties or chemical events.

## STRUCTURE DESCRIPTION AT AN ATOM LEVEL

In the development of nonempirical structure descriptors of molecular structure, attention has focused on indexes representing the whole molecule. The term "nonempirical" in this context means that the index is directly computable from molecular structure, that is, from the connection table. From the earliest developments by Wiener[4-6] and Platt,[7] indexes have been based on atoms or bonds, but these features are subsumed into whole molecule indexes. These quantities, applied to alkanes, have been exclusively topological in nature. Molecular connectivity indexes, developed by Kier and Hall[2,3] from the Randić-alkane branching index,[8] are currently the most widely used.[9-11] The valence $\chi$ indexes of molecular connectivity include electronic characterization of the atom valence state as well as topological factors. These indexes, as well as others, have been applied to the analysis of physicochemical and biological properties (QSAR) quite successfully.[9-11] To avoid a semantic issue concerning whether these indexes should be called topological, we shall refer to the broader category as indexes from chemical graph theory.

The basic elements of chemical graph theory are counts of structure features in the molecular skeleton. Graph theoretical indexes are developed from counting features such as atoms, bonds, rings, pairs of adjacent bonds, and so forth. The objective is the development of graph invariants, indexes which represent the molecular structure and which are independent of the manner in which the graph is numbered. Indexes from graph theory have become the basis of QSAR expressions whose statistics are often impressive.[9-11] However, these indexes are summations over all the designated features in the molecular skeleton. Such overall sums tend to obscure atomic level information so that the atom or fragment level implications are not readily apparent, except for the $^0\chi$ and $^0\chi^v$ indexes of Kier and Hall, which are defined as an atom index.[2,3] In earlier work, Hall and Kier suggested an approach to revealing such important information by decomposing whole molecule indexes into subgraphs.[12] In this approach, subgraphs for a limited portion of the molecule were substituted for the whole molecule indexes. There was little loss in the statistical quality of the model, but there was a gain in understanding

0095-2338/91/1631-0076$02.50/0 © 1991 American Chemical Society

of the part of the molecule most important to activity. This method is useful but somewhat cumbersome, and more directly methods would be helpful.

The development of indexes from chemical graph theory and models built upon them has brought about a quantitative expression of the basic concept that structurally similar molecules behave in similar fashion. It is expected, and experience has indicated, that small variations in graph theory indexes relate to variations in activities and properties. However, especially in the biological milieu, individual atoms or localized molecular regions, such as a pharmacaphore, may play a major or dominant role. Whole molecule indexes may tend to obscure this fact and prevent clear identification of such atomic features. There is, therefore, a need to explore the possibility of more directly applicable atom-level indexes.

Chemical graph theory uses the chemical graph (hydrogen-suppressed skeleton) for generation of atom-level structure indexes. Starting with molecular connectivity methodology, Kier and Hall[13,14] developed indexes to relate to partial charge in alkanes. Three atom-level indexes were calculated by partitioning whole molecule $\chi$ indexes (${}^0\chi$, ${}^1\chi$, and ${}^2\chi$) into atom contributions. This work was extended by Kier[15] to an expression for the uniqueness of an atom in a molecule and further developed by Hall and Kier into a description of the topological equivalence of atoms.[16] Other approaches have been reported, including summed atom values to give the molecular ID numbers by Randić,[17] the torsion topological descriptors of Nilakantan,[18] the electron-topologic approach of Bersuker,[19] the topological electronic index of Kaliszan,[20] a vertex topological index by Klopman,[21,22] and the development of atomic contributions for physicochemical properties by Crippen and Ghose.[23]

In this work we present the development of a new and general approach to atom-level indexes which encode both electronic and topological aspects of the atom in the molecular skeleton.

## INTERACTIONS OF ATOMS IN A FIELD

In this study, each atom in a molecule is considered to reside in a field composed of every other atom in that molecule. It is the impact of other atoms on a given atom that is the basis for the information which defines the state of the atom within the molecular skeleton. The result of these interactions is the modification of the intrinsic state of the given atom to produce the bonded state of the atom within the context of the whole molecule. If these influences are defined within a model combining electronic and topological attributes of an atom and its field, the resulting index is called the electrotopological state of that atom.

In this study the objective is to develop expressions for the information present in each field and to assemble this into an electrotopological-state index associated with each atom of the molecule.

## QUANTITATIVE EXPRESSION OF THE FIELD

**The General Model.** In our view, an expression for the graph field effect upon a given atom includes three basic components. First, the intrinsic topological and electronic state of the atom is encoded. Second, the effect of the field in its influence on an atom is quantified. The third component is the dependence of the interaction on the graph separation of atoms in the molecule. Each of these aspects will be developed separately.

**Intrinsic Atom Value.** First, second-quantum-level atoms (e.g., B, C, N, O, F) will be considered. The intrinsic state of a skeletal atom must possess a dual character, expressing both electronic and topological information. In this approach, it is structure information which is represented, not experimental properties of the molecule. First, consider as the electronic factor in this model the valence-state electronegativity. The count of $\pi$ and lone-pair electrons associated with a skeletal atom is used to encode this property. This count of non-$\sigma$ electrons has been shown to correlate well with the valence-state electronegativity of (second-quantum-level) covalently bound atoms.[2,3,15] This non-$\sigma$ electron count for a given atom or group is equal to $\delta^v - \delta$ where $\delta^v$ is the count of valence electrons in the skeleton ($Z^v - h$) and $\delta$ is the count of $\sigma$ electrons in the skeleton ($\sigma - h$). For a skeletal atom, $Z^v$ is the number of valence electrons, $\sigma$ is the count of electrons in $\sigma$ orbitals, and $h$ is the number of bonded hydrogen atoms.[24] For example, in $NH_2$, $\delta^v = 3$ $(5 - 2)$; $\delta = 1$ $(3 - 2)$; and $\delta^v - \delta = 2$.

Table I. Intrinsic-State Values

| atom (skeletal hydride group) | intrinsic-state value[a] |
|---|---|
| $\rangle C \langle$ | 1.250 |
| $\rangle CH-$ | 1.333 |
| $-CH_2-$ | 1.500 |
| $\rangle C=$ | 1.667 |
| $-CH_3$, $=CH-$, $\rangle N-$ | 2.000 |
| $\equiv C-$, $-NH-$ | 2.500 |
| $=CH_2$, $=N-$ | 3.000 |
| $-O-$ | 3.500 |
| $\equiv CH$, $-NH_2$ | 4.000 |
| $=NH$ | 5.000 |
| $\equiv N$, $-OH$ | 6.000 |
| $=O$ | 7.000 |
| $-F$ | 8.000 |
| $-Cl$ | 4.111 |
| $-Br$ | 2.750 |
| $-I$ | 2.120 |
| $=S$ | 3.667 |
| $-SH$ | 3.222 |
| $-S-$ | 1.833 |

[a] Calculated from eq 2.

For the topological attribute to be encoded into the atom intrinsic value, consider the atom's relative location within the molecule or what might be described as its relative degree of mantle-atom or buried-atom status. For example, the methyl groups of neopentane are mantle atoms, whereas the central carbon is a buried atom; a tertiary amino nitrogen is more topologically surrounded than is a terminal primary amino group. To encode this information, use is made of the count of nearest neighbors in the skeleton, $\delta$, as a measure of the degree of mantle-atom status. Skeletal atoms for which $\delta$ = 3 or 4 are relatively buried within the molecule whereas terminal atoms, for which $\delta = 1$, tend to lie on the surface or mantle of the molecule.

It is proposed here that the atom intrinsic value be a function of $\delta^v - \delta$ and $\delta$. To begin the development, consider the ratio of these two quantities for the intrinsic quantity of interest: $(\delta^v - \delta)/\delta$. However, in this simple ratio, all of the hydrides of an atom in the same valence state have identical values, such as the C(sp$^3$) hydrides ($CH_3-$, $-CH_2-$, $\rangle CH-$, $\rangle C \langle$). It is most desirable that these groups not have degenerate intrinsic values because their roles in physicochemical or biological processes are not identical. It is observed that the addition of a constant to the numerator breaks this degeneracy. Accordingly, the value of $\delta^v - \delta$ is scaled by adding one: $(\delta^v - \delta + 1)/\delta$.

This relation may be simplified by adding one to the whole expression and reducing the terms to get the expression which is adopted provisionally for the atom intrinsic value, $I$:

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

A list of $I$ values for second-quantum-level atoms is shown in Table I. The treatment of higher row atoms follows in the next section.

**Higher Quantum-Level Atoms.** The relation developed above for intrinsic values includes an expression related to the electronegativity of second-quantum-level atoms, approximated by the quantity $\delta^v - \delta$. In a series of atoms with constant $\delta$ value, such as the series -O-, -NH-, -$CH_2$- ($\delta = 2$), the variation in the $\delta^v$ value encodes the relative electronegativity of the group. However, in a family, such as the halogens, there is a constant value of $\delta^v - \delta$ (when $\delta^v$ is taken as $Z^v - h$) but not constant electronegativity. To reflect adequately the differences in electronegativity among family members for the intrinsic value (eq 1), it is necessary to characterize the relative electronegativities in the series. For this relation, an important quantity from the quantum mechanical treatment of atoms, the principal quantum number, $N$, is adopted.

For purposes of this development, it is proposed that the ratio of the squares of the principal quantum number relative to the second quantum level ($N = 2$) be used as a modifier of the $\delta^v$ value in eq 1:

$$I = [(2/N)^2\delta^v + 1]/\delta \tag{2}$$

This is the relation used for all atoms. All the $I$ values in Table I are computed with eq 2, including several atoms and groups of higher quantum levels shown at the bottom of Table I.

**Graphical Expression for Effect of Atom on Each Atom.** The influence on atom $i$ by all the other atoms in the field may be dissected into a summation of the relationships of all atom pairs, $i...j$. The total effect is a summation of the effect of all atom pairs.

To develop a quantitative relation, we assume that the field has a perturbing effect on the intrinsic atom value $I$; this perturbation is assumed to be some function of the difference in intrinsic values $I_i$ and $I_j$. Thus, the perturbation of atom $i$ from the interaction with atom $j$ is given as a function of the difference in intrinsic values:

$$\Delta I_i = \mathbf{f}(I_i - I_j) \tag{3}$$

This equation is adopted as a general expression for the perturbation of a given atom by another atom.

Further, it is expected that the influence of atom $j$ on atom $i$ is less for widely separated atoms; thus, graph separation must be a factor in the final expression. We modify eq 3 with a function of $r_{ij}$, the count of atoms in the shortest path between atoms $i$ and $j$, including both $i$ and $j$. (Note that $r_{ij}$ is not the graph distance, $d_{ij}$, the count of edges between $i$ and $j$.) Current work is based upon the assumption that the interaction effect decreases with the square of the graph separation, $r_{ij}^2$. The graph field effects are summed, and a relation is obtained for the influence of the field upon the intrinsic value of an atom:

$$\Delta I_i = \sum_{j=1}^{N} (I_i - I_j)/r_{ij}^2 \tag{4}$$

As a result, this influence on $I$ leads to an expression for the electrotopological state of atom $i$, $S_i$:

$$S_i = I_i + \Delta I_i \tag{5}$$

The quantity $S_i$ is called the electrotopological state of atom $i$ or, simply, the E-state of an atom.

## EXAMPLES

**Sample Calculation.** Computation of the E-state index for each atom begins with the hydrogen-suppressed graph, which is essentially the molecular skeleton, as illustrated in Figures 1–3. Input to a computer program is accomplished with the molecule connection table which specifies the atom type, the number of bonded hydrogens, and the identification of each connected atom. Such input can be done in several formats and may be facilitated with a graphic form of input.[25]

To illustrate the steps in the procedure and to make explicit the contribution from each pair of atoms, it is useful to examine the components of the E-state for each atom in a molecule. For this example consider alanine. The intrinsic $I$ value for each atom is computed from eq 2. In Table II, the various pair contributions are given in matrix form. Note that this E-state matrix is antisymmetric. This summation across row $i$ gives the field effect perturbation on each atom, $\Delta I_i$. The state values for each atom are calculated and are shown at the bottom of the table: $S_i = I_i + \Delta I_i$.

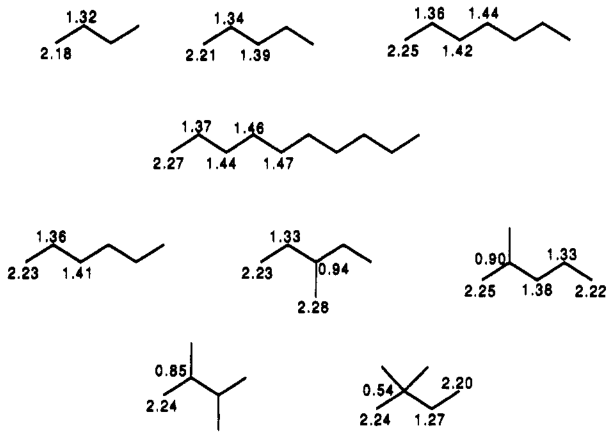


**Figure 1.** Alkane graphs with computed electrotopological-state values which illustrate the effects of chain lengthening and chain branching.

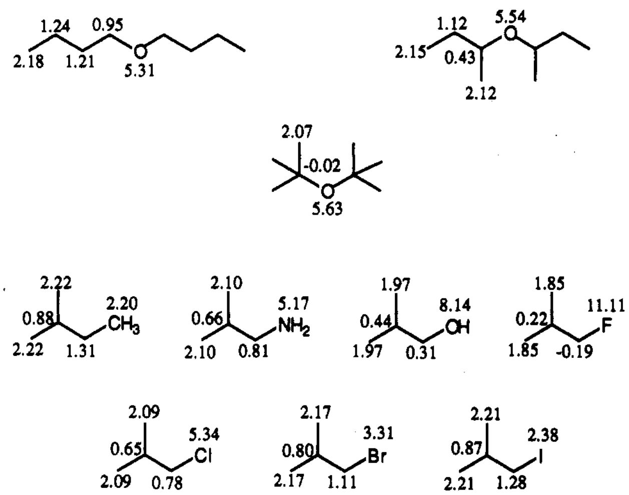
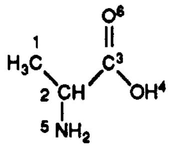


**Figure 2.** Molecular skeletons containing heteroatoms with computed electrotopological-state values which illustrate effects of branching and heteroatom substitution.

**Results of Calculations.** Several normal alkanes and the hexane isomers are shown in Figure 1 along with the E-state values for each skeletal atom. For the unbranched alkanes, the terminal methyl group increases in E-state value with chain lengthening. Also the central portion of the molecule builds in E-state value but at a lower level than the more exposed terminal methyl group. For very long chain normal alkanes, the terminal methyl group tends to a value of 2.33 and the central atoms to 1.50. The penultimate methylene group is decreased in E-state value below its intrinsic value (1.50) whereas the middle methylenes approach the intrinsic value for long normal alkanes. The terminal methyl is increased above its intrinsic value (2.00). For the range of carbon atoms from a primary to a quaternary environment, the mantle-atom status declines. The E-state values parallel this structure attribute. Terminal methyl groups have E-state values around 2.23–2.28 whereas the quaternary carbon value is 0.54. Methylene and methine carbons possess values in between these two extremes.

Figure 2 presents molecule graphs with heteroatoms. The greater electron richness of oxygen results in significantly

**Table II.** Electrotopological-State Calculations for Alanine

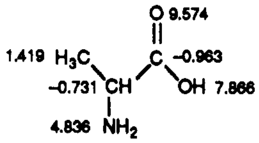


| intrinsic values | intrinsic values | intrinsic values |
|---|---|---|
| $I(1) = 2.000$<br>$I(2) = 1.333$ | $I(3) = 1.667$<br>$I(4) = 6.000$ | $I(5) = 4.000$<br>$I(6) = 7.000$ |

$(I_i - I_j)/r_{ij}^2$ Matrix

| | *j* | | | | | | |
|---|---|---|---|---|---|---|---|
| *i* | 1 | 2 | 3 | 4 | 5 | 6 | $\Delta I =$ row sum |
| 1 | 0.0 | 0.1667 | 0.0370 | −0.2500 | −0.2222 | −0.3125 | −0.5810 |
| 2 | −0.1667 | 0.0 | −0.083 | −0.5185 | −0.6667 | −0.6296 | −2.0648 |
| 3 | −0.0370 | 0.0833 | 0.0 | −1.0833 | −0.2593 | −1.3333 | −2.6296 |
| 4 | 0.2500 | 0.5185 | 1.0833 | 0.0 | 0.1250 | −0.1111 | 1.8657 |
| 5 | 0.2222 | 0.6667 | 0.2593 | −0.1250 | 0.0 | −0.1875 | 0.8356 |
| 6 | 0.3125 | 0.6296 | 1.3333 | 0.1111 | 0.1875 | 0.0 | 2.5741 |
| | | | | | | | 0.0000 |

$$S_i = I_i + \Delta I_i$$



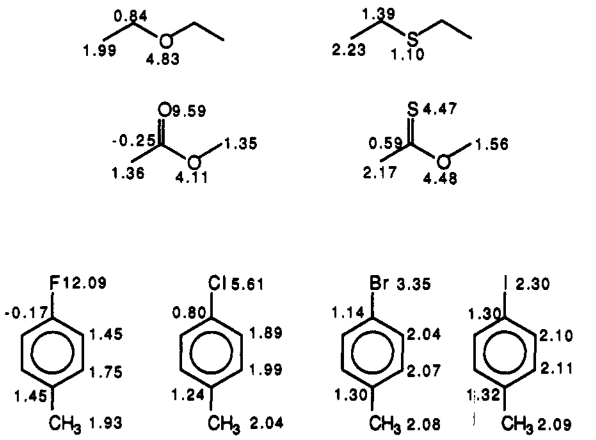


**Figure 3.** Molecular skeletons which illustrate the change in electrotopological state with the variation of higher row heteroatoms.

**Table III.** Electrotopological-State Values for Alkyl Ethers with Computed Oxygen Partial Charges and $^{17}O$ NMR Chemical Shift

| obs | compound | $q(\text{–O–})^a$ | $S(\text{–O–})^b$ | $^{17}O$ $\delta^c$ | calc$^d$ | res$^e$ |
|---|---|---|---|---|---|---|
| 1 | dimethyl ether | −0.2971 | 4.20 | −52.2 | −52.9 | 0.7 |
| 2 | ethylmethyl- | −0.3060 | 4.54 | −22.5 | −21.4 | −1.1 |
| 3 | isopropylmethyl- | −0.3133 | 4.75 | −2.0 | −2.0 | −0.0 |
| 4 | *t*-butylmethyl- | −0.3166 | 4.94 | 8.5 | 15.6 | −7.1 |
| 5 | diethyl- | −0.3149 | 4.83 | 6.5 | 5.4 | 1.1 |
| 6 | isopropylethyl- | −0.3220 | 5.04 | 28.0 | 24.9 | 3.1 |
| 7 | *t*-butylethyl | −0.3253 | 5.23 | 40.5 | 24.4 | −1.9 |
| 8 | diisopropyl- | −0.3301 | 5.25 | 52.5 | 44.3 | 8.2 |
| 9 | *t*-butylisopropyl- | −0.3334 | 5.44 | 62.5 | 61.9 | 0.6 |
| 10 | di-*t*-butyl | −0.3341 | 5.63 | 76.0 | 79.5 | −3.5 |

[a] Oxygen partial charge computed by the STO-3G method, ref 28. [b] Electrotopological-state value for ether oxygen. [c] Measured $^{17}O$ NMR chemical shift ref 27. [d] Chemical shift obtained from regression of experiment chemical shift on the E-state value for oxygen, eq 6. [e] Res = obs − calc.

larger E-state values than for carbon. Also as the branching in the vicinity of the ether oxygen increases, its E-state value also increases. Here the large electronegativity of the oxygen has greater impact than its increasing buried status. Conversely, the atom attached to oxygen has a decreasing E-state value as branching increases at that atom. However, in the series of alkyl amines (hexyl, dipropyl, and triethyl), the nitrogen atom-state value actually decreases, reflecting the surrounded topology of the tertiary nitrogen. The effect of electronegativity is also revealed in the series of isobutyl derivatives. In the order of increasing electronegativity, the heteroatom E-state value increases. The other atoms decrease in the expected manner. The effect of decreasing electronegativity in the halogen series is also clearly shown in the molecules in the bottom of the figure.

Figure 3 shows E-state values for oxygen and sulfur compounds. The sulfur field impact is less than that of oxygen. Also shown are halogen-substituted toluenes. The effect of changing the halogen on the ipso and para atoms of benzene rings is mirrored by the E-state values. The halogens have decreasing E-state values with higher quantum numbers, and the consequent effect on the other atoms can be seen.

**Relation of E-State Values to NMR Chemical Shift.** In the formulation of the E-state, electronic and topological effects are combined. The E-state index for an atom is based upon the valence-state electronegativity, derived from the π and lone-pair electron count, in addition to a quantitative expression of atom topology due to its immediate surroundings. Accordingly, we would expect that this index, the E-state value of an atom, would bear some relationship to a physical measurement which is dependent upon electron density and molecular topology. We have shown in earlier papers that the $^{17}O$ NMR chemical shift for both a set of alkyl ethers and a set of aldehydes and ketones is highly correlated to the E-state value for oxygen.[26]

We have also developed a relation between chemical shift and computed partial charge on the carbonyl oxygen. Our intent here is not a comparison between E-state values and computed partial charges. The interest lies in comparison of the quality of models based on an E-state index and models based on a quantity computed by a quantum mechanical approach. Fliszar has computed the oxygen partial charge using STO-3G ab initio wave functions.[27,28]

**Table IV.** Alkyl Aldehydes and Ketones Electrotopological-State Values, Oxygen Partial Charges, and $^{17}O$ NMR Chemical Shifts

| obs | compound | $q(O)^a$ | $S(=O)^b$ | $\delta^c$ | calc$^d$ | res$^e$ |
|---|---|---|---|---|---|---|
| 1 | $CH_3CHO$ | −0.2289 | 8.806 | 592.0 | 590.0 | 2.0 |
| 2 | $C_2H_5CHO$ | −0.2296 | 9.174 | 579.5 | 579.7 | −0.2 |
| 3 | $i$-$C_3H_7CHO$ | −0.2301 | 9.505 | 574.5 | 570.5 | 4.0 |
| 4 | $(CH_3)_2CO$ | −0.2667 | 9.444 | 569.0 | 572.2 | −3.2 |
| 5 | $CH_3COC_2H_5$ | −0.2695 | 9.813 | 557.5 | 562.0 | −4.5 |
| 6 | $CH_3CO$-$i$-$C_3H_7$ | −0.2681 | 10.144 | 557.0 | 552.8 | 4.2 |
| 7 | $(C_2H_5)_2CO$ | −0.2721 | 10.181 | 547.0 | 551.8 | −4.8 |
| 8 | $C_2H_5CO$-$i$-$C_3H_7$ | −0.2707 | 10.512 | 543.5 | 542.6 | 0.9 |
| 9 | $(i$-$C_3H_7)_2CO$ | −0.2738 | 10.843 | 535.0 | 533.4 | 1.6 |

[a] Oxygen partial charge computed by the STO-3G method, ref 28. [b] Electrotopological-state value for the carbonyl oxygen. [c] $^{17}O$ NMR chemical shift, ref 27. [d] Calculated chemical shift from eq 8. [e] Res = obs − calc.

For the ethers the partial charges, $q$(-O-), are given in Table III along with the electrotopological-state values for the ether oxygen, $S$(-O-).

Table IV contains comparable information for the carbonyl compounds: the partial charges, $q(=O)$, and the E-state values, $S(=O)$.

For the ethers, the statistical results are as follows:

$$\delta = 92.564(\pm 3.331)S(\text{-O-}) - 441.652(\pm 16.660)$$
$$r = 0.995,\ s = 4.3,\ F = 772,\ n = 10 \quad (6)$$

$$\delta = -3292.9(\pm 94.7)q(\text{-O-}) - 1031.6(\pm 30.3)$$
$$r = 0.997,\ s = 3.4,\ F = 1208,\ n = 10 \quad (7)$$

where $S$(-O-) is the ether oxygen E-state index and $q$(-O-) is the ether oxygen partial charge computed by Fliszar. For the aldehydes and ketones, the statistical results are as follows:

$$\delta = -27.77(\pm 1.98)S(=O) + 834.48(\pm 19.47)$$
$$r = 0.983,\ s = 3.67,\ F = 197,\ n = 9 \quad (8)$$

$$\delta = 791.01(\pm 168.91)q(=O) + 764.65(\pm 43.47)$$
$$r = 0.871,\ s = 9.75,\ F = 22,\ n = 9 \quad (9)$$

where $S(=O)$ is the E-state value for the carbonyl oxygen and $q(=O)$ is the partial charge computed by Fliszar. The standard error on the regression coefficients is given in parentheses after each coefficient. The observed, calculated, and residual chemical shifts are given in Tables III and IV.

These relations indicate a high quality of relationship between the computed E-state index for the oxygen atoms and the corresponding $^{17}O$ NMR chemical shift. These models, based on E-state relations, are statistically equivalent to the models based on STO-3G ab initio partial charges for the ethers and superior to those for the carbonyl compounds. Clearly the E-state index encodes important and relevant electronic and topological information for these ether and carbonyl oxygen atoms.

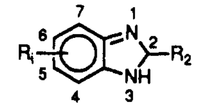
**Inhibition of Lee Strain of Flu Virus by Benzimidazoles.** Another illustration of the significance of the E-state values is their use in a biological investigation. Tamm et al.[29] produced a set of data for the inhibition of flu virus by benzimidazoles as given in Table V. We have computed the E-state values for the nine skeletal atoms in the structurally invariant part of the molecules, the benzimidazole ring system, as given in Table V. Because of the tautomerism involving the two nitrogen atoms, the intrinsic values of the two nitrogen atoms have been set equal to their average value, $I(N) = 2.75$.

Through standard multiple linear regression techniques it was found that two atom E-state values gave an excellent correlation equation, as follows:

$$pK_i = 4.267(\pm 0.436)S(N_{1,3}) + 0.786(\pm 0.193)S(C_2) - 15.995(\pm 2.044)$$
$$r = 0.956,\ s = 0.16,\ F = 63,\ n = 15 \quad (10)$$

The variable $S(N_{1,3})$ is the average of the E-state values of the two nitrogen atoms, $N_1$ and $N_3$; $S(C_2)$ is the E-state value for the 2-position carbon atom. The standard error on the regression coefficients are given in parentheses. The observed, calculated, and residual activities are also given in Table V.

**Table V.** Inhibition of Lee Strain Flu Virus by Benzimidazoles with the Electrotopological State Values



| obs | substituents | $S(N_{1,3})^a$ | $S(C_2)^b$ | $pK_i^c$ | calc$^d$ | res$^e$ |
|---|---|---|---|---|---|---|
| 1 | benzimidazole | 4.008 | 1.574 | 2.14 | 2.35 | −0.21 |
| 2 | 2-methyl- | 4.202 | 0.848 | 2.51 | 2.60 | −0.09 |
| 3 | 5-methyl- | 4.059 | 1.587 | 2.72 | 2.57 | 0.15 |
| 4 | 5,6-dimethyl- | 4.110 | 1.601 | 2.72 | 2.80 | −0.08 |
| 5 | 4,6-dimethyl- | 4.139 | 1.608 | 2.82 | 2.93 | −0.11 |
| 6 | 2,5-dimethyl- | 4.253 | 0.852 | 2.89 | 2.82 | 0.07 |
| 7 | 4,5-dimethyl- | 4.139 | 1.608 | 2.96 | 2.93 | 0.03 |
| 8 | 2,5,6-trimethyl- | 4.304 | 0.856 | 3.05 | 3.04 | 0.01 |
| 9 | 2,4,5-trimethyl- | 4.333 | 0.860 | 3.20 | 3.17 | 0.03 |
| 10 | 5,6-diethyl- | 4.192 | 1.629 | 3.39 | 3.17 | 0.22 |
| 11 | 2-propyl-5-methyl- | 4.442 | 0.975 | 3.60 | 3.73 | −0.13 |
| 12 | 2,4,5,6,7-pentamethyl- | 4.464 | 0.871 | 3.66 | 3.74 | −0.08 |
| 13 | 2-ethyl-5-methyl- | 4.371 | 0.940 | 3.74 | 3.40 | 0.34 |
| 14 | 2-butyl-5-methyl- | 4.490 | 0.993 | 3.77 | 3.95 | −0.18 |
| 15 | 2-isopropyl-5-methyl- | 4.452 | 0.945 | 3.77 | 3.75 | 0.02 |

[a] Equation variable taken as the average of the E-state indexes for the two nitrogen atoms in the benzimidazole moiety. See text. [b] E-state index for the carbon atom in the 2-position. [c] Experimental value for the inhibition constant. See ref 29. [d] Values computed for inhibition from eq 10. [e] Res = $pK_i$ − calc.

**Table VI.** Summary of the Leave-One-Out Method on the Benzimidazole Flu Inhibition Data

| quantity$^a$ | from full regression$^b$ | from leave-one-out method$^c$ |
|---|---|---|
| $S(N_{1,3})$ | 4.28 (0.44) | 4.29 (0.15) |
| $S(C_2)$ | 0.786 (0.19) | 0.790 (0.05) |
| intercept | −15.99 (2.04) | −16.08 (0.68) |
| average residual | 0.116 | 0.145 |

[a] See text and Table V for definitions. [b] From eq 10. [c] Mean and standard deviation from the 15 regressions in which each observation was deleted, one at a time. See text.

In this straightforward manner, it is shown that the nitrogens and the 2-position of the benzimidazole play an important role in the biological activity. These results are consistent with earlier results which indicate that substitution at the 2-position is more important to high activity than substitution on other positions.[12]

Inspection of the relation between activity and the E-state value for the nitrogens indicates two groupings of molecules, depending on whether the molecule is substituted in the 2-position. Although the correlation between the activity and $S(N_{1,3})$ is significant ($r = 0.794$), there are clearly two distinct groups of compounds when regression is based only on $S(N_{1,3})$: those substituted at the 2-position and those not. The high quality of the two-variable correlation clearly indicates that the structure information contained in $S(C_2)$ is appropriate in its effect to unite the two groups of molecules in the final equation. For purposes of drug design, the medicinal chemist can focus attention on the 1,2,3-region of the benzimidazole moiety in order to enhance activity. Since the coefficients of both E-state indexes are positive, substitutions which increase their values should enhance activity.

To investigate the robust quality of eq 10, a leave-one-out method has been applied. Each observation was deleted from the data set, one at a time; the regression analysis was run on the diminished set; the activity of the deleted observation was predicted, and its difference from the observed value computed (residual). Table VI summarizes the results of this investigation as the mean and standard deviation of the regression coefficients along with the average residual of the predicted (deleted) observations. This information is listed under "leave-one-out method" in Table VI; the statistics for the single full regression are also given for comparison. The residuals on predicted (deleted) observations range from −0.20 to 0.34, essentially the same as in the full regression (see Table V). As can be clearly seen in Table VI, the coefficients are very stable and the predicted values are very good, leading to errors on predicted values which are quite acceptable. This leave-one-out method simulates the predictive quality of the regression equation in a satisfactory manner.

## DISCUSSION

The electrotopological-state index of an atom unifies in a single index both an electronic and a topological description. These structure attributes are encoded directly from counts of skeletal features which include electronic and topological information. Constructed in this way, the E-state indexes are computable directly from a connection table, without the need of adjustable parameters.

The atom intrinsic value $I$ is derived from the count of $\pi$ and lone-pair electrons, $\delta^v - \delta$. It has been shown that this count is related to the valence-state electronegativity of the skeletal atom. Thus, the difference in intrinsic values, $\Delta I$, for a pair of skeletal atoms encodes both electronic and topological attributes which arise from electronegativity difference and skeletal connectivity. Derived from this electronegativity difference, the E-state value for an atom is related in part but not limited to the concept of atomic partial charge.

The E-state index also depends on the molecular topology of the skeleton, information which enters the formalism in two ways. The atom intrinsic value $I$ contains in the denominator the count of skeletal neighbors of the atom, $\delta$, a measure of the local topology. This neighbor count is the number of $\sigma$ bonds in the immediate skeletal neighborhood, that is, an expression of the electronic connection of that atom to the rest of the molecule. Further, in the $\Delta I$ expression, the topology of the whole molecule enters through the graph separation, $r$. The electronegativity difference is diminished by the $r^2$ factor where $r$ is the count of atoms in the skeleton between pairs of atoms (including both atoms). Hence, there is a strong proximity-related topological dependence in the E-state formulation.

The E-state value of an atom, $S$, may be viewed as a sum of effects due to skeletal atoms at various distances:

$$S_i = \underset{\substack{\text{intrinsic}\\ \text{value}}}{I_i} + \underset{\substack{\text{bonded}\\ \text{contributions}}}{\sum(I_i - I_j)/4} + \underset{\substack{\text{nonbonded}\\ \text{contributions}}}{\sum(I_i - I_j)/r^2} \qquad (11)$$

The atom $i$ state value is the intrinsic value $I_i$ modified by the perturbing terms. The $\Delta I$ terms may be positive or negative, depending on the relative magnitudes of $I_i$ and $I_j$; $S$ values for atoms of larger $I$ values are increased through the interaction, and $S$ values for smaller $I$ values are decreased. In part, the effect of atoms bonded to atom $i$, $\Delta I/4$, (those $\alpha$ to it) produce a varying degree of polarity or charge buildup, depending upon the magnitude of $\Delta I$ for those atom interactions. The effect of atoms $\beta$, $\gamma$, $\delta$, etc. to atom $i$ are nonbonded interactions, diminished by the factor $r^2$: 9, 16, 25, etc.

The influence term $\Delta I_i$ can be either positive or negative. When $I_i > I_j$, the corresponding contribution to $\Delta I_i$ is positive. When most of the terms in the series are positive, then $\Delta I_i > 0$. Hence, terminal atoms of high electronegativity tend to have $S$ values much greater than their initial intrinsic value $I$. Less electronegative atoms, which are buried in the skeleton, tend to decrease in value (from the intrinsic value) and may acquire negative E-state values.

Description of some of the E-state structure information is presented in this paper with a series of related molecular structures. The variation of E-state values with skeletal branching agrees with common organic intuition. Likewise, the variation of the E-state value among different heteroatoms and in different topological settings is satisfying with respect to experience with such effects as polarity and inductive effects.

It is shown that the correlation of E-state values with $^{17}O$ NMR chemical shift is quite significant for a set of carbonyl compounds which includes both aldehydes and ketones. It is observed that as the E-state value increases, the chemical shift decreases. A similar high-quality relation was observed for a set of 10 ethers.[26] For the carbonyl compounds, partial charges computed from optimized STO-3G wave functions produced a much less significant correlation. These studies indicate that a high level of electrotopological information is encoded in the E-state index. Although improved ab initio computations may well lead to higher quality correlations, it is sufficient here to note the high quality of the E-state correlation. It is not intended that the E-state values are a more simply estimated atomic partial charge.

It is to be emphasized that the E-state indexes are said to be nonempirical, that is, they are computable directly from the molecular connection table. They are derived from counts of electrons within the hybridization model of covalent binding and from the adjacency relations in the hydrogen-suppressed graph. The E-state indexes are the first atom-level indexes which combine into a single index both electronic structure and molecular topology by using the same metric, electron counts. The calculation is simple and straightforward, requiring only the element type, the hybrid state, and the connection table of atoms in a molecular skeleton.[25]

The potential value of this index is apparent when the need for atom-centered, nonempirical structure descriptors is realized. There now exists the possibility to examine submolecular features to discover contributions toward intermolecular effects among biologically important molecules in such phenomena as protein binding and receptor interactions. To the significant arsenal of whole molecule indexes, valuable in their own sphere of utility, is now added a new set of tools to expedite investigation of molecular mechanism and rational design of molecules at the atomic and fragment level.

## ACKNOWLEDGMENT

We express appreciation for valuable discussions with Jack W. Frazer and for support of this work from Sterling Drug, Inc., Malvern, PA, and Allied Signal Corporation, Engineered Materials Research, Des Plaines, IL.

**Registry No.** $CH_3CHO$, 75-07-0; $C_2H_5CHO$, 123-38-6; $i$-$C_3H_7CHO$, 78-84-2; $(CH_3)_2CO$, 67-64-1; $CH_3COC_2H_5$, 78-93-3; $CH_3CO$-$i$-$C_3H_7$, 563-80-4; $(C_2H_5)_2CO$, 96-22-0; $C_2H_5CO$-$i$-$C_3H_7$, 565-69-5; ($i$-$C_3H_7)_2CO$, 565-80-0; benzimidazole, 51-17-2; 2-methylbenzimidazole, 615-15-6; 5-methylbenzimidazole, 614-97-1; 5,6-dimethylbenzimidazole, 582-60-5; 4,6-dimethylbenzimidazole, 69557-54-6; 2,5-dimethylbenzimidazole, 1792-41-2; 4,5-dimethylbenzimidazole, 69557-55-7; 2,5,6-trimethylbenzimidazole, 3363-56-2; 2,4,5-trimethylbenzimidazole, 62192-75-0; 5,6-diethylbenzimidazole, 69557-56-8; 2-propyl-5-methylbenzimidazole, 5473-06-3; 2,4,5,6,7-pentamethylbenzimidazole, 69700-34-1; 2-ethyl-5-methylbenzimidazole, 30411-81-5; 2-butyl-5-methylbenzimidazole, 7479-04-1; 2-isopropyl-5-methylbenzimidazole, 4887-90-5; alanine, 56-41-7; dimethyl ether, 115-10-6; ethyl methyl ether, 540-67-0; isopropyl methyl ether, 598-53-8; $t$-butyl methyl ether, 1634-04-4; diethyl ether, 60-29-7; isopropyl ethyl ether, 625-54-7; $t$-butyl ethyl ether, 637-92-3; diisopropyl ether, 108-20-3; $t$-butyl isopropyl ether, 17348-59-3; di-$t$-butyl ether, 6163-66-2.

## REFERENCES AND NOTES

[1] Elliel, E. L. *Stereochemistry of Carbon Compounds*; McGraw-Hill: New York, 1982; p 1.

(2) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(3) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; John Wiley: London, 1986.

(4) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17, 2636.

(5) Wiener, H. *J. Chem. Phys.* **1947**, *15*, 766.

(6) Wiener, H. *J. Phys. Chem.* **1948**, *52*, 425, 1082.

(7) Platt, J. R. *J. Phys. Chem.* **1952**, 56, 328.

(8) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.

(9) Trinajstić, N. *Chemical Graph Theory*; CRC: Boca Raton, FL, 1983; Vols. I and II.

(10) Rouvray, D. H. *Sci. Am.* **1986**, *255*, 40.

(11) Sabljić, A.; Trinajstić, N. *Acta Pharm. Jugosl.* **1981**, *31*, 189.

(12) Hall, L. H.; Kier, L. B. *J. Pharm. Sci.* **1978**, *67*, 1743.

(13) Hall, L. H. in *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990; pp 202-236.

(14) Kier, L. B.; Hall, L. H. *Tetrahedron* **1977**, *33*, 1953.

(15) Kier, L. B. *Quant. Struct.-Act. Relat.* **1987**, *65*, 8.

(16) Hall, L. H.; Kier, L. B. *Quant. Struct.-Act. Relat.* (in press).

(17) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.

(18) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82.

(19) Bersuker, I. B.; Dimoglia, A. S.; Gorbachov, M. Yu. In *QSAR in Drug Design and Toxicology*; Hadzi, D., Jerman-Blazic, B., Eds.; Elsevier: Amsterdam, 1987; p 43.

(20) Kaliszan, R. *Chromatography* **1987**, *29*, 19.

(21) Klopman, G.; Raychandhury, C. *J. Comp. Chem.* **1988**, *9*, 232.

(22) Klopman, G.; Raychandhury, C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 12.

(23) Ghose, A. K.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21.

(24) Kier, L. B.; Hall, L. H. *J. Pharm. Sci.* **1981**, *70*, 583.

(25) The computations on all data presented in this manuscript were performed with a new version of MOLCONN2 which will be available in the near future from L. H. Hall.

(26) Kier, L. B.; Hall, L. H. *Pharm. Res.* **1990**, *7*, 801.

(27) Delseth, C.; Kintzinger, J.-P. *Helv. Chim. Acta* **1976**, *59*, 466, 1411.

(28) Fliszar, S. *Charge Distributions and Chemical Effects*; Springer-Verlag: New York, 1983; p 63.

(29) Tamm, I.; Folkers, K.; Shunk, C. H.; Horofall, F. L. *J. Exp. Med.* **1953**, *98*, 245.

# A Comment on Nomenclature and the Unsaturated Bond

D. J. POLTON*

Department of Computer Science, University of Hull, Hull HU6 7RX, England

Received July 4, 1990

The rules for naming compounds containing both double and triple bonds are unnecessarily complicated by the seniority given to the double bond over the triple bond. Although it is realized that the situation cannot be changed without great upheaval, it is hoped that by drawing attention to an anomalous situation which has been perpetrated throughout the world of chemical notation as well as nomenclature, any newly devised system will be a little better by avoiding it.

One of the curiosities of most, if not all, forms of chemical nomenclature and notation is the order of seniority allotted, for the purpose of enumeration, to the degree of unsaturation of carbon–carbon bonds. Double bonds are considered the most senior form of bond, although they are an intermediate form between the fully hydrogenated (single) bond and the fully unsaturated (triple) bond.

In reality, there should only be two possible orders of seniority of bonds, that of increasing or decreasing hydrogenation between the two atoms. Other orders, where the single or the triple bond comes between the others, do not seem to be logical. It could be said that the double bond is of more frequent occurrence than the triple bond and should therefore take precedence over it, but the bond by far most frequently seen in structures is single. And as for alphabetization, -en comes between -an and -yn.

It is therefore logical that if the double bond is considered to be senior to the triple bond then the single bond must be senior to the double bond. As in general nomenclature and notation, if the former is true then the single bond should be senior to the double bond.

Consider the simplest case of propene:

$$CH_3{-}CH_2{=}CH_2$$

Taking the single bond to be senior, the enumeration should be

$$\overset{1}{C}H_3{-}\overset{2}{C}H_2{=}\overset{3}{C}H_2$$

But in IUPAC nomenclature[1] it is, of course

$$\overset{3}{C}H_3{-}\overset{2}{C}H_2{=}\overset{1}{C}H_2$$

That is, the double bond is senior to the single bond.

Therefore the triple bond should logically be senior to the double bond.

In IUPAC nomenclature the endings -ane, -ene, and -yne are given to compounds containing single bonds only, single and double bonds only, and triple with any other bond type, respectively. Structures with both double and triple bonds have the "en" in a position inferior to the "yne", as in -enyne or -adienyne. This implies that the most senior bond type is triple, as its ending overrides the presence of double bonds. Quite probably this order stems from alphabetization, -a, -e, -y, which is incidentally the order of increasing unsaturation.

Thus, although consideration of locants for a series of similar features before consideration of what those features are specifically is very important in any hierarchical scheme, the seniority given to the double bond for the purpose of enumeration does seem to be anomalous.

Further complication ensues on enumeration of a chain containing both double and triple bonds. IUPAC rule A-3.3 requires that the lower (at the first point of difference) numbering set be assigned to the positions of unsaturation, whether double or triple bonds. If both a double and a triple bond occupy the same position with respect to the ends of the chain, then the double bond is taken as prior. Single bonds are ignored.

The following examples of IUPAC numbering illustrate this:

$$\overset{1}{C}H{\equiv}\overset{2}{C}{-}\overset{3}{C}H{=}\overset{4}{C}H{-}\overset{5}{C}H_3 \qquad \text{(I)}$$

$$\overset{1}{C}H_2{=}\overset{2}{C}H{-}\overset{3}{C}{\equiv}\overset{4}{C}{-}\overset{5}{C}H_3 \qquad \text{(II)}$$

$$\overset{4}{C}H{\equiv}\overset{3}{C}{-}\overset{2}{C}H{=}\overset{1}{C}H_2 \qquad \text{(III)}$$

In I and II, 1,3 being lower than 2,4, the prior consideration is that terminal bond having the highest degree of unsaturation.

* Present address: Rosefarm; Denne Manor Lane, Shottenden, Canterbury, Kent CT4 8JJ, England.

0095-2338/91/1631-0082$02.50/0 © 1991 American Chemical Society