BCSJ

Structural and Physicochemical Interpretation of GT-STAF Information Theory-Based Indices

Stephen J. Barigye,*[1,2] Yovani Marrero-Ponce,[2,3,4,5] Jure Zupan,[6] Facundo Pérez-Giménez,[4] and Matheus P. Freitas[1]

[1]Departamento de Química, Universidade Federal de Lavras, UFLA, Caixa Postal 3037, 37200-000 Lavras, MG, Brazil

[2]Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Universidad Central "Martha Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba

[3]Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, E-46071, València, Spain

[4]Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain

[5]Facultad de Química Farmacéutica, Universidad de Cartagena, Cartagena de Indias, Bolívar, Colombia

[6]Laboratory of Chemometrics, National Institute of Chemistry, Ljubljana, Hajdrihova 19, 1000 Ljubljana Slovenia

E-mail: sjbarigye@gmail.com
Received: February 9, 2014; Accepted: August 29, 2014; Web Released: September 8, 2014

The underlying structural and physicochemical interpretation of the recently defined information indices (denominated as GT-STAF indices) is examined, with the aim of gaining greater insight on the codified chemical information. It is found that these indices are related with molecular symmetry in the context of the defined molecular "fragment" model. Moreover, these indices are sensitive to structural differences, demonstrating gradual changes consistent with modifications in the molecular structure. A principal component analysis reveals that the GT-STAF indices generally codify conformational, physicochemical, and thermodynamic properties of amino acids. A study with aniline derivatives demonstrates that the GT-STAF indices do not directly correlate with the ionization constant (pK_a); but rather require multivariate contributions to yield correlations comparable with univariate models for quantum chemical parameters, suggesting that the former codify some other form of electronic information orthogonal to the latter. Finally, an evaluation of atomic contributions to the molecular hydrophobicity in furylethylenes demonstrates that the GT-STAF approach generally approximates to chemical properties quite well.

The emergence of theoretical molecular descriptors pioneered by the seminal work of Harold Wiener,[1] presents a ground breaking mathematical approach in characterizing molecular structures. This seemingly oversimplified approach has surprisingly proved to codify important structural information and ensuing parameters have been successfully used in various structure–activity relationship studies. However, an interjection to the keen interest in these theoretical parameters has been the interrogative: "what do these numerical values mean, or what attribute (or chemical phenomenon) do they represent?" Although it is commendable that a molecular descriptor (MD) be interpretable,[2] it is not imperative that these parameters have meaning, or be explainable in terms of well-known parameters (or dimensions), as some of these are in fact MDs as well, rationalized only in the considered models.

In addition, the success of theoretical MDs in correlating with a wide range of properties suggests that these contain numerous finer features buried in single parameters. It is therefore intuitive that one approach may not suffice to give an adequate approximation to the real chemical image captured by the molecular parameters and may thus require combined contributions from different chemical models. In fact, several manuscripts by different authors may be cited in the literature devoted to the interpretation of the same family of indices, for example, the molecular connectivity indices.[3–10] However, these are just a handful relative to the volumes of manuscripts in which theoretical molecular parameters are proposed.[11] This reflects the complexity rather than a rebuttal to the challenge to offer interpretations to MDs using mainstream chemical frameworks.

In previous reports,[12–16] the theoretical aspects of the GT-STAF (acronym for graph theoretical Thermodynamic state functions) information indices, based on the insight of a chemical structure as a communication system were presented. Accordingly, Shannon's entropy, mutual information, conditional and joint entropy-based information indexes (IFIs) were defined for binary, triple, and quadruple dimensional systems. Also diverse criteria for generating molecular "fragment" models, as information sources, were considered with the aim of obtaining different perspectives of the chemical structure and thus achieve greater approximation to chemical reality.[14] These models are classified in three main groups,

that is, *graph-theoretic models* (among which are: connected subgraphs (CS), walks of length k (K), terminal paths (TP), vertex paths incidence (VP), quantum (Q) and Sachs (S) subgraphs), *fingerprint-based models* (which include: MACCs (MA), substructure (SS) and E-state (ES) keysets), and *magnitude-based models* [atomic hydrophobicity (ALOP) and (AMR) refractivity magnitudes]. A series of studies with the GT-STAF IFIs yielded satisfactory results, generally superior to those reported in the literature with other methods.[12–16] Nonetheless, greater comprehension of the intrinsic information codified with these indices would enhance their practical utility in chemoinformatic tasks.

Consequently, the present manuscript is dedicated to the investigation of structural and physicochemical interpretation of the GT-STAF IFIs in their different extensions, using insights from diverse approaches. First, a brief recapitulation of the theoretical aspects of the proposed IFIs will be given.

Theoretical Scaffold

The GT-STAF Information Indices. Shannon's Entropy as Information Index: Consider as an information source a set of molecular "fragments" S that describes a molecular structure. It follows that the constituent atoms (vertices) of the molecular "fragments" possess different participation frequencies in the set S. Therefore, from these participation frequencies, a probability distribution function (p.d.f) is constructed. Applying Shannon's fundamental equation

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i \tag{1}$$

where p_i is the probability associated with vertex v_i and n is the number of vertices that constitute molecular graph G, yields Shannon's entropy-based information index (IFI) for the analyzed molecular structure.

Other molecular structure entropy parameters like negentropy and standardized Shannon's entropy[11,17] could be applied to the p.d.f obtained for a given chemical source, yielding the corresponding IFIs.

Channel Coding Theorem in Information Index Derivations: In addition to eq 1, other entropic measures like mutual information index (*MI*, eq 2), joint entropy index (*JE*, eq 3) and conditional entropy index (*CE*, eq 4) could also be computed obtaining corresponding molecular IFIs:

$$H(X;Y) = MI(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

$$H(X,Y) = JE(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \tag{3}$$

$$H(Y|X) = CE(Y|X) = H(X,Y) - H(X) \tag{4}$$

where the elements p(x, y) form the matrix P:

$$P(X,Y) = \left\{ p(x,y):\, p(x,y) = f(x,y)/f_T | x \neq y \wedge f_T = \sum_{x=1}^{n} \sum_{y=1}^{n} f_{(x,y)},\ x = y \right\} \tag{4a}$$

Note that when $p(x,y) = p(x)p(y)$, H(X;Y) = 0 and when $p(x,y) = p(x)$, H(X;Y) = H(X).

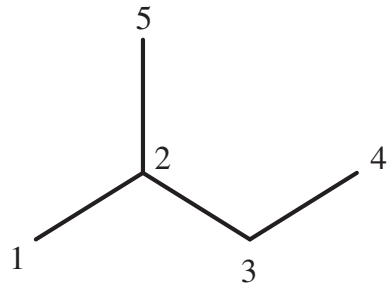


Figure 1. The chemical structure of a molecule of isopentane (the numbers correspond to the labels that are assigned to the non-hydrogen atoms (vertices)).

Let us take as a simple example a molecular graph of isopentane (Figure 1), where the numbers correspond to the labels that are assigned to the carbon atoms (vertices) in the molecular structure.

In this illustration the connected subgraphs model, based on graph-theoretical concept of subgraph orders is used as a data source.

Accordingly, for the molecular graph (G) in Figure 1, the connected subgraphs obtained for different orders based on the atomic relations are:

Order 0: C_1, C_2, C_3, C_4, C_5
Order 1: C_1–C_2, C_2–C_3, C_3–C_4, C_2–C_5
Order 2: C_1–C_2–C_3, C_1–C_2–C_5, C_2–C_3–C_4, C_3–C_2–C_5
Order 3: C_1–C_2–C_3–C_4, C_5–C_2–C_3–C_4, C_1–C_2–(C_5)–C_3
Order 4: C_1–C_2–(C_5)–C_3–C_4

We may be interested in assigning code words to source symbols (vertices) using a coding-tree scheme based on the incidence of vertices in the molecular "fragments" (subgraphs), forming an m-length binary code words, where m is the number of molecular "fragments" ($m = 17$ in this case). In this scheme, code words are assigned to vertices using successive choices between 0 and 1 at each branch in the coding-tree structure. Given a set of subgraphs, $S = \{s_g |\ 1 \leq g \leq s_m\}$, the code word for v_i is sequentially assigned:

1, if v_i is included in s_g, where $1 \leq g \leq s_m$
0, otherwise

For the chemical source (set of 17 subgraphs) generated for orders 0–4 above (shown in Figure 2A in a random manner to mimic normal text), the corresponding fixed length (17 bit) code words for the vertices v_i would therefore be:

C_1 11000110000010101
C_2 10110110111011101
C_3 10100011011101101
C_4 10100001001000110
C_5 10011010011010000

Definitely the interest here is not code optimality but rather dissimilarity of the different vertex code words, if applicable, and indistinctive code words (not uniquely decodable) are not "penalized" but rather considered informational about the structural similarity of the compared vertices.

Suppose the code word for vertex C_3 is transmitted along an ideal noiseless channel,[18–20] it is certain that at the receiver's end the same code word will be received. However, for a noisy channel,[18–20] several possibilities exist: code word sequence for

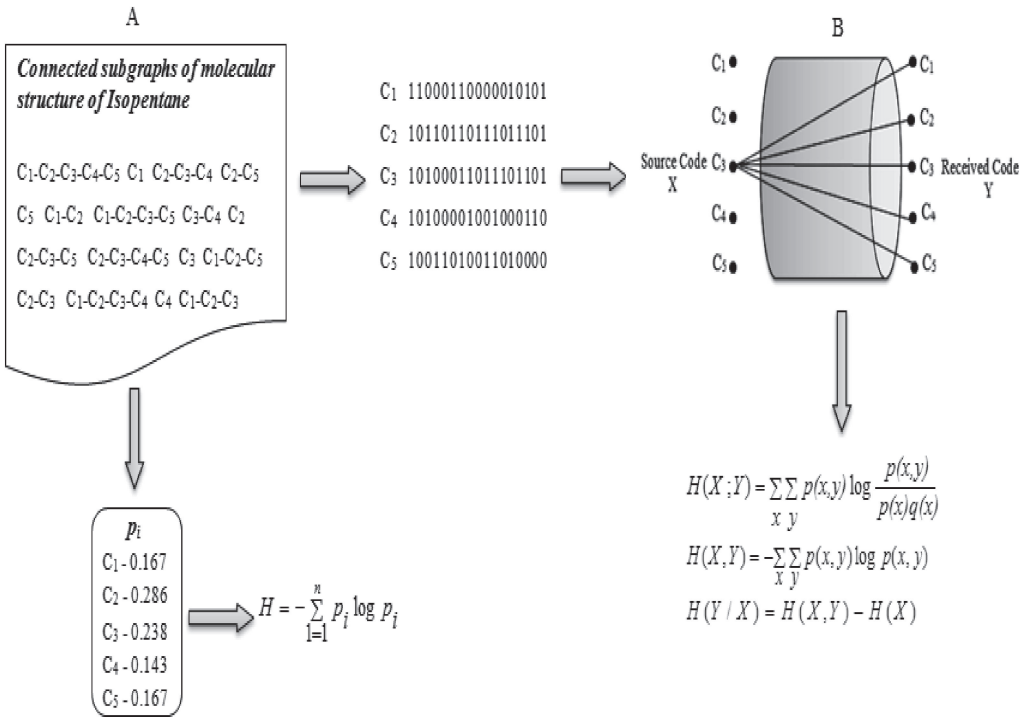


Figure 2. A) Illustration of chemical source entropy computation for the molecular structure of isopentane. The chemical source is comprised of connected subgraphs (for more information see Refs. 1–3). B) Schematic representation of the relations between inputs and outputs in a noisy channel. Note that the *MI*, *JE*, and *CE* analysis is carried out for each of the input codes with respect to the output codes.

vertex C_1 is received instead of the one corresponding to C_3, C_2 instead of C_3, C_4 instead of C_3, or C_5 instead of C_3 (see illustration in Figure 2B). This means that the received message is not necessarily the same as that sent out by the transmitter. The mutual information (MI) for vertex code words for v_x and v_y, $H(v_x, v_y)$ gives a measure of the true information content at the receiver's end.

For all vertex code word pairs $({}^{c}v_x, {}^{c}v_y)$, the respective mutual frequencies are computed as a tally of 1 bit length correspondences between the code words. These mutual frequencies $f(x,y)$ are subsequently used to compute the joint probabilities $p(x,y)$ for 1 bit length "sequences" using eq 4a and thus a joint p.d.f $P(X,Y)$ is formed. Note that there is no real distinction between vertices x and y; this designation is hypothetical.

Let us illustrate the computation of the MI, CE, and JE IFIs using eqs 1, 2, 3, 4, and 4a. For operational convenience, mutual frequencies and corresponding joint probabilities are represented by frequency and joint probability matrices, denoted by **F** and **P**, respectively, as shown below:

$$\mathrm{F} = \begin{bmatrix} 7 & 6 & 4 & 2 & 3 \\ 6 & 12 & 8 & 4 & 6 \\ 4 & 8 & 10 & 5 & 4 \\ 2 & 4 & 5 & 6 & 2 \\ 3 & 6 & 4 & 2 & 7 \end{bmatrix} \qquad \mathrm{P} = \begin{bmatrix} 0.167 & 0.143 & 0.095 & 0.048 & 0.071 \\ 0.143 & 0.286 & 0.190 & 0.095 & 0.143 \\ 0.095 & 0.190 & 0.238 & 0.119 & 0.095 \\ 0.048 & 0.095 & 0.119 & 0.143 & 0.048 \\ 0.071 & 0.143 & 0.095 & 0.048 & 0.167 \end{bmatrix}$$

$$MI(X,Y) = 3.001 \text{ bits} \qquad JE(X,Y) = 6.566 \text{ bits.}$$

The $CE(Y/X)$ for G is obtained by substituting the values for $H(X) = JE(X,X)$ and $JE(X,Y)$ in eq 4 and by the chain rule, $CE(Y/X) = 4.294$ bits.

This approach is extended to consider information coding for communication systems with three and four source dimensions, respectively, and corresponding applications to molecular structure codification are derived, see Ref. 13.

Moreover, in the codification of molecular structure information, it is desirable that the indices used permit discrimination of isomeric structures. Consequently, in Ref. 12, schemes for codification of heteroatoms and unsaturated bonds are formulated to achieve greater applicability of the GT-STAF approach in the characterization of structural information. Also generalizations of the summation operator as the global characterization of the vertex code word entropies are discussed.

The different approaches (or models) aimed at providing structural and physicochemical interpretations to the proposed IFIs will now be discussed, with the hope of achieving better comprehension of the information codified by these parameters,

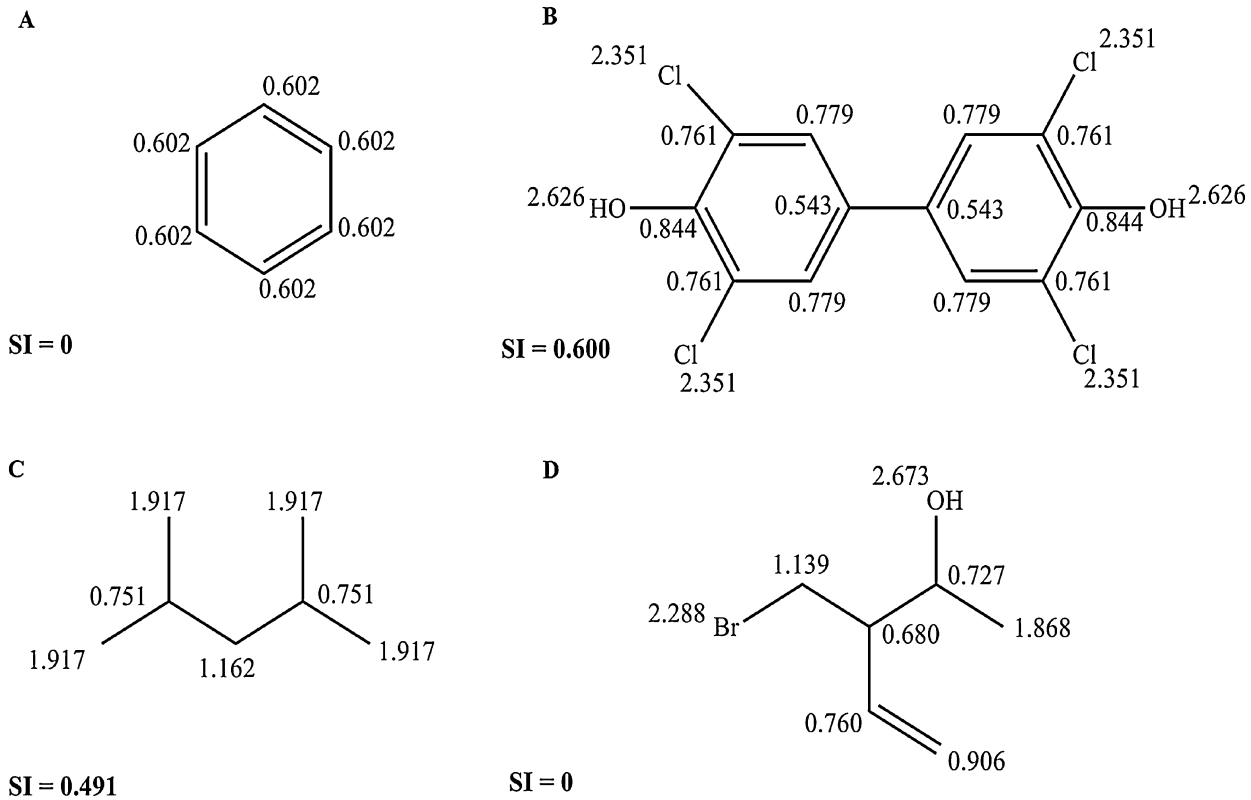


Figure 3. Atomic mutual information values for molecular structures **A**, **B**, **C**, and **D**, computed using the connected subgraphs molecular "fragment" model and Pauling's electronegativity scale.

as well as to appreciate their usability in chemoinformatic tasks.

Results and Discussion

Molecular Symmetry Perspective. The outlook of a chemical structure as a communication system permits similarity-based analyses to be performed and conveys intuitive understanding on the nature of the molecular structure. Note that for this case both the source originator (input) and destination (output) belong to the same molecule.

When the code word for vertex v_i is transmitted along a communication channel and an "identical code word" (in this context refers to same bit sequence or Hamming weight,[20] i.e. number of 1 bits in a given code word) belonging to another vertex v_j is received at the output terminal, this suggests that these vertices are similar (i.e. possess same identity and/vicinity). On the other hand, completely "dissimilar code words" represent vertices of different nature, in structural terms (equivalent to transmission along useless communication channels). In between these two extrema, we may have a code word for v_i to some extent similar to that of v_j but with some distortions (modifications), typical of a nonideal or noisy communication channel. In this case, knowledge on the degree of similarity for the analyzed code words would reflect on the extent of structural equivalence of the considered vertices.

In order to assess the trueness of these information-theoretic inferences on the nature of vertices, *MI* values for each vertex with respect to the rest of vertices for molecular structures **A**–**D** were computed, using the connected subgraphs model as the data source and Pauling's electronegativity as the weighting scale, to permit the discrimination of unsaturated bonds and heteroatoms (Figure 3). Note that *SE*, *CE*, or *JE* values of vertices could be used as well and these would yield the same corollaries.

If the vertex *MI* values for the molecular structures **A**–**D** are examined, one feature stands out, vertices with similar environs in a molecular structure, have identical *MI* values, while vertices with dissimilar vicinities have different *MI* values. Certainly, permutations for such vertices are structure preserving. It is thus natural that these entropy values are related to molecular symmetry on a bi-dimensional scale. Determining the proportion of structure-preserving vertex-permutations in a molecular structure yields a measure of molecular symmetry. Characterizing the statistical distribution of vertex entropy values offers a suitable measure of molecular symmetry, but caution should be taken for the possible subjectivity of such measure with respect to the number of entropy values, that is, molecular size. Accordingly the normalized or standardized Shannon's entropy (*sSE*) is used,[11] defined as:

$$sSE = \frac{-\sum_{i=1}^{G} \frac{\rho(g_i)}{N} \log_2 \frac{\rho(g_i)}{N}}{\log_2 N} \quad (5)$$

where $\rho(g_i)$ is the cardinality of equivalence class g_i, and N, is the number of vertices of molecular structure. The *sSE*, which will be denominated the *symmetry index*, *SI*, has the following characteristics:

$0 \leq SI \leq 1$

$SI = 0$ represent perfectly symmetry

$SI = 1$ represents maximum asymmetry, from a topological perspective

These attributes have been suggested as imperative for ideal symmetry measures.[11] Consequently, structure **A**, characterized by identical vertex *SE* values, has $SI = 0$, and thus perfectly symmetrical, while for structure **D**, with distinct vertex *SE* values, yields $SI = 1$, representing maximum asymmetry.

Symmetry plays a vital role in the quantum-mechanical understanding of atomic and molecular states, NMR spectra and various physicochemical properties, such as entropy, dipole moment, surface tension, etc.[21–25] Thus the GT-STAF approach should yield good correlations for these properties. With the aim of evaluating the validity of these deductions, we compare the atomic entropy values computed for oxygen atoms in a set of ethers and carbonyl compounds with their corresponding ^{17}O nuclear magnetic resonance (NMR) chemical shifts,[26] using single variable regression equations (note that no weighting scheme was used). As anticipated, good correlations are observed between the atomic entropy values and the chemical shifts, for ethers (eq 6 and Table 1) and carbonyl compounds (eq 7, for details see Table SI1, Supporting Information), respectively.

Ethers

$$\Delta = -105.64\ (\pm 7.50) + 51.40\ (\pm 2.95) CE_{0\text{–}1} \qquad (6)$$

$$R^2 = 0.97,\ s = 6.77,\ F = 304.16,\ n = 10$$

Carbonyls

$$\Delta = 627.61\ (\pm 5.99) - 34.87\ (\pm 3.06) CE_{1} \qquad (7)$$

$$R^2 = 0.95,\ s = 4.49,\ F = 129.48,\ n = 9$$

Besides yielding good correlations, other interesting tendencies are observed. In the first place, atomic entropy values computed from low order (0–1) "fragments" yield better correlations than the ones computed from the entire model of molecular "fragments" (for details see Table SI2, Supporting Information). This result suggests less relevance for distant methyl group contributions to the chemical shifts, which is consistent with spectroscopic understanding. For example, it is known that the inductive effect of substituent groups on proton shifts influences the separation only up to 3 bonds as it is rapidly buffered with increase in distance.

Another interesting observation from this study is that, compared to *SE*, *JE*, and *MI* values much better correlations are obtained with *CE* values. From information theory, *CE*, also known as equivocation, measures the level of dissimilarity or ambiguity between the sent and received messages. Likewise, for atomic *CE* measures the degree of dissimilarity between the codes for the considered atom and those for atoms in its vicinity. On the other hand, chemical shifts reflect the alteration of the environment of an atom in a molecule due to electronic and steric influences. An attempt to rationalize the superior correlations for *CE* with ^{17}O chemical shifts from an information theoretic perspective suggests that chemical shifts are thus related to a measure of "ambiguity" between the environment of an atom in a "promolecule" (comprised of atoms of identical vicinities) and the actual compound in whose vicinity it is analyzed. This is in fact not far from nuclear magnetic resonance intuition. Chemical shifts express the amount by which proton resonance is shifted with respect to a standard reference substance, whose protons are the most shielded.

Finally, given that chemical shifts mirror the milieu of an atom in a molecule due to electronic and topological influences, the good correlations for the atom entropy measures and this property suggest that these indices represent a unified approach effective in encoding both electronic and steric attributes of atoms in molecules.

Table 1. Experimental and Calculated Values for ^{17}O NMR Chemical Shifts in Ethers

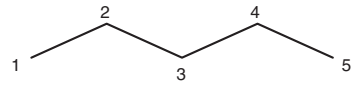
ID	Compound	$CE_{0\text{–}1}$	^{17}OΔ	Calculated
1	Dimethyl	0.99	−52.2	−54.78
2	Ethyl methyl	1.59	−22.5	−23.84
3	Isopropyl methyl	2.09	−2	2
4	*t*-Butyl methyl	2.52	8.5	23.73
5	Diethyl	2.11	6.5	2.97
6	Isopropyl ethyl	2.54	28	24.96
7	*t*-Butyl ethyl	2.90	40.5	43.62
8	Diisopropyl	2.91	52.5	43.99
9	*t*-Butyl isopropyl	3.23	62.5	60.32
10	Di-*t*-butyl	3.51	76	74.84

Influence of Structural Changes on Atomic Entropy Values. One of the desirable properties of novel *MD*s is that these possess progressive and logical variations on gradual structural changes. An analysis of these trends may provide greater understanding of the information codified by proposed *MD*s.[26] Consequently, in this section the tendencies of the GT-STAF IFIs in different structural variations are evaluated. The entropy computations for subsection: **Molecular Symmetry Perspective** to subsection: **Amino Acid Properties and their Relation to GT-STAF Indices** are carried out using the connected subgraphs molecular "fragment" source originator, and for subsection: **Evaluation of Relationship between Experimental pK_a Values of Substituted Anilines and GT-STAF Indices**, two molecular "fragment" models are selected for each group, i.e. CS and S for graph theoretic models, MA and SS keysets for fingerprint models; and ALOGP and AMR sets for magnitude-based models.

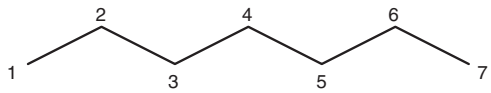
Chain Lengthening in Alkanes: As it can be observed in Table 2, lengthening of the carbon chain is accompanied by gradual reductions in the terminal and inner vertex SEs. On the other hand, a contrasting trend is observed with *CE*, in the sense that with chain lengthening, the terminal and inner vertex *SE*s increase. As for *MI*, while the inner vertex *MI*s progressively increase, no definite trend is traceable for the terminal *MI*. Terminal *JE* increases from pentane to heptane and then it gradually decreases for the subsequent terminal *JE* values. In the case of inner *JE* values; these increase up to heptane, after which the addition of an ethyl group to the carbon skeleton, adds one inner vertex, from each periphery inwards, whose *JE* value reduces.

Branching in the Carbon Skeleton: For *SE*, the effect of branching leads to an increase in the entropy value at the branching point, with more buried vertices having higher *SE* values (for details see Table SI3, Supporting Information). An opposite trend is observed for *MI*, *CE*, and *JE*, with more buried vertices characterized by declines in their respective entropy values. Entropy values for vertices adjacent to

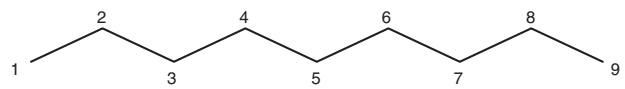
Table 2. Influence of Chain Lengthening



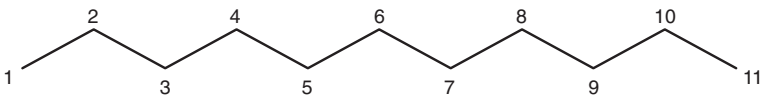
	1	**2**	**3**
SE	0.401	0.487	0.504
MI	0.580	0.688	0.700
CE	2.062	1.730	1.511
JE	2.149	2.417	2.428



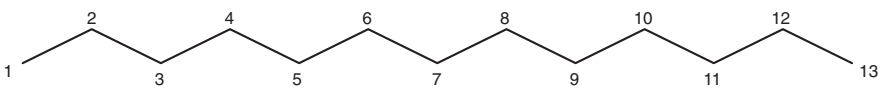
	1	**2**	**3**	**4**
SE	0.299	0.401	0.444	0.456
MI	0.683	0.914	0.989	1.005
CE	2.817	2.753	2.388	2.225
JE	2.610	3.196	3.328	3.344



	1	**2**	**3**	**4**	**5**
SE	0.216	0.318	0.380	0.411	0.420
MI	0.650	0.941	1.078	1.148	1.168
CE	3.227	3.772	3.637	3.352	3.230
JE	2.130	3.104	3.720	3.933	3.983



	1	**2**	**3**	**4**	**5**	**6**
SE	0.172	0.260	0.316	0.354	0.379	0.387
MI	0.661	1.013	1.179	1.225	1.229	1.238
CE	3.464	4.367	4.611	4.600	4.485	4.415
JE	1.805	2.736	3.383	3.880	4.282	4.388



	1	**2**	**3**	**4**	**5**	**6**	**7**
SE	0.144	0.221	0.272	0.307	0.331	0.345	0.352
MI	0.654	1.048	1.270	1.358	1.349	1.291	1.252
CE	3.665	4.816	5.280	5.464	5.547	5.601	5.622
JE	1.573	2.441	3.074	3.573	3.977	4.294	4.498

branching points are equally influenced by increase in ramifications of the carbon skeleton in a similar manner. This result suggests the existence of a relationship between the topological arrangement of atoms and their entropy values. Hence properties closely related to the topology of molecular structures should be properly codified by these indices.

Unsaturated Bonds in Carbon Skeleton: In the GT-STAF approach, a weighting scheme is adapted to codify unsaturated bonds in order to discriminate sp/sp^2 carbon atoms from corresponding sp^3 carbons. It is thus logical that only the entropy values for involved vertices are altered with respect to the saturated analogs (for details see Table SI4, Supporting Information).

It follows that, unsaturation leads to a decrease in the entropy values, proportional to the vertex degree of the involved atoms, that is, there is more decrease for sp than sp^2 hybridized carbon atoms, as well as for terminal (sp/sp^2) than inner (sp/sp^2) carbon atoms. Therefore unsaturated bonds "invert" the roles of the vertices that constitute the unsaturated bonds. In molecular sciences, in order to gain better compre-

hension of the influence of unsaturated bonds on a molecular structure, concepts like bond length, electron density, rigidity, etc., are almost classical. For example, from bond length, that is, the mean distance in time between two nuclei of two atoms bonded together, it is known that the Csp^2–Csp^2 bond is shorter than Csp^3–Csp^3. Although the adapted scheme was simply aimed at achieving discrimination for unsaturated bonds, it would be plausible that the entropy values have some sort of relationship with properties like bond length. Following this notion, the unsaturated bond entropies were computed (as the average entropy for the atom pairs that constitutes the unsaturated bonds, parallel to the bond length concept) and their correlations with the bond length found (see Table SI5, in Supporting Information).

The correlation coefficients, *R*, obtained for *SE*, *MI*, *CE*, and *JE* with the bond length were 0.995, 0.998, 0.867, and 0.995, respectively. Thus, other than for *CE*, the application of the weighting scheme to *SE*, *MI*, and *JE* achieves good relations with the bond length, suggesting that not only is discrimination of unsaturated bonds achieved, but this scheme permits codifying important structural information consistent with the introduction of unsaturated bonds to the carbon skeleton.

Introduction of Heteroatom in Carbon Skeleton: Given that standard graph-theoretic molecular "fragment" models do not possess the capacity of discriminating heteroatoms from carbon atoms, a scheme was adapted which assigns weights to the atoms in the molecular structure to award greater applicability to the IFIs derived from these models. These weights are atomic properties such as: van der Waals volume, Pauling's electronegativity, etc. Therefore for functional group isomers characterized by changes in the heteroatom type, the effect of the introduction of the heteroatom is largely dependent on the trend offered by the weighting scheme, in the sense that, if a property decreases across and/or down the periodic table, corresponding reductions will be achieved in the entropy values, and vice versa.

Cyclicity and Aromaticity: Most drugs have as their nuclei, cyclic or aromatic structures, which make these structures of particular interest in medicinal chemistry. It is thus desirable that proposed MDs adequately discriminate isomeric cyclic and/or aromatic structures, or more specifically unsaturated bonds, from this class of compounds. This study is aimed at evaluating to what extent the GT-STAF IFIs adequately discriminate unsaturated bonds in diverse cyclic or aromatic vicinities. Comparisons with linear analogs are made as well. Table 3 shows the *SE*, *MI*, *CE*, and *JE* values calculated for the selected molecular "fragment" models.

Firstly, while cyclization (hex-1-ene to cyclohexane and but-1-ene to cyclobutane), is accompanied by an increase in all entropic values for graph theoretic models, the ALOGP-based entropic values depict the contrary, decreasing for all entropic values. As for AMR-based IFIs, cyclization causes reduction in *SE* and *CE*, whereas increases are observed for *MI* and *JE* IFIs. In the case of the fingerprint-based IFIs, the small size of the considered molecules means that these will have very few (or none) of the fingerprints activated and thus they do not yield meaningful values especially for *MI*, *CE*, and *JE* in some of the cases. However, it could be noted that for MA fingerprint-based models show an increase *SE* values on cyclization.

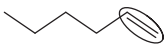
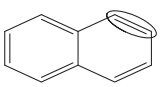
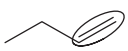
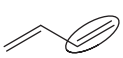
A shift from saturated to unsaturated bonds causes a decrease in all graph-theoretic entropy values, as well as *SE* values for the MA's model. An opposite trend is observed for ALOGP entropy values, increasing on moving from saturated to unsaturated bonds. In the case of AMR models, an increase is observed for *SE* and *CE*, while a decrease for *MI* and *JE* values is obtained. On the other hand, in graph-theoretic events, no differences in entropy values are observed for unsaturated bonds in conjugated (or aromatic) and unconjugated systems, neither does the inclusion of heteroatoms (see pyridine) affect these entropic values. As for the MA-based model, while conjugation does not yield differences in entropy values for unsaturated bonds, aromaticity produces a decrease in *SE* values. The SS-based model yields no value for unconjugated bonds in cyclic carbon skeletons, although a decrease is observed for *SE* values on moving from nonaromatic to aromatic conjugated bonds. A comparison of entropy values for ALOGP and AMR reveals contrasting trends: while conjugation is accompanied with a decrease in all the entropy values for ALOGP, in the case of AMR, conjugation causes an increase the corresponding entropic measures. Aromaticity causes an increase in *SE*, *MI*, and *JE* values for ALOGP and a decrease for *CE* in the same model. As for AMR, a decrease is observed for *SE* and *CE* values while an increase is obtained for *MI* and *JE* values. The inclusion of a heteroatom in the aromatic system causes a decrease in *SE* values for MA and SS-based models. In the ALOGP and AMR models, *SE* and *CE* values increase while *MI* and *JE* values decrease.

Finally, in graph-theoretic models, unsaturated bonds in naphthalene possess lower *SE* values and higher *MI*, *CE*, and *JE* values than in benzene. Likewise, unsaturated bonds in naphthalene for MA and SS models possess lower *SE* than in benzene. As for ALOGP models, all entropic values in naphthalene are lower than those for benzene. The same tendency is observed for AMR models with the exception of *CE* values.

Here more than revealing the changes in entropy values for unsaturated bonds in different vicinities, it is known that the reactions that involve unsaturated bonds largely depend on the accessibility of the π electrons, which is in turn largely influenced by the vicinity of these electrons. Therefore parameters that depict gradual changes (increases or reductions) for unsaturated bonds in different vicinities, should codify important structural information inherently useful in the prediction of properties inherent to the nature of unsaturated bonds.

General Trends for Terminal and Inner Atoms Entropies. As may have been perceived in Table 3, *SE*, *MI*, *CE*, and *JE* computations using the CS source originator algorithm assign more weight to inner vertices than peripheral ones, thus suggesting that in the codification of molecular properties, greater role is placed on the contribution of inner vertex entropies than the terminal ones. On the other hand, if the weighting scheme for discriminating unsaturated bonds or heteroatoms is applied to saturated hydrocarbons, the role of the terminal vertex entropies with respect to the inner vertex entropies is "reversed," conveying a dominant role to the former. The question as to which scheme is the most optimal has no clear cut answer, as this depends on the property considered. Nonetheless, it has been demonstrated that for bond

Table 3. Influence of Cyclicity and Aromaticity

Compounds	Index[a)]	CS[b)]	S[c)]	MA[d)]	SS[e)]	ALOGP[f)]	AMR[g)]
	*SE*	8.144	7.685	9.354	3.742	3.080	5.238
	MI	11.659	11.125	0.000	0.000	0.540	0.832
	CE	36.090	41.828	0.000	0.000	6.043	8.543
	JE	42.213	55.140	0.000	0.000	0.561	0.862
	*SE*	9.672	9.672	9.672	0.000	1.526	4.836
	MI	15.907	12.813	0.000	0.000	0.403	2.510
	CE	50.431	45.672	0.000	0.000	0.806	5.020
	JE	68.398	68.712	0.000	0.000	0.403	2.510
	*SE*	6.448	6.448	7.128	0.000	2.252	5.580
	MI	10.605	8.542	0.000	0.000	0.451	1.489
	CE	33.621	30.448	0.000	0.000	3.158	10.421
	JE	45.599	45.808	23.347	0.000	0.451	1.489
	*SE*	6.448	6.448	6.777	0.000	2.425	5.089
	MI	10.605	8.542	0.000	0.000	0.919	2.348
	CE	33.621	30.448	0.000	0.000	2.297	5.869
	JE	45.599	45.808	11.673	0.000	0.919	2.348
	*SE*	6.448	6.448	6.777	7.483	2.394	5.120
	MI	10.605	8.542	0.000	0.000	0.906	2.371
	CE	33.621	30.448	0.000	0.000	2.264	5.927
	JE	45.599	45.808	11.673	0.000	0.906	2.371
	*SE*	6.448	6.448	6.448	6.448	2.659	4.744
	MI	10.605	8.542	0.000	0.000	1.169	2.438
	CE	33.621	30.448	0.000	0.000	1.169	2.438
	JE	45.599	45.808	0.000	0.000	1.169	2.438
	*SE*	4.663	4.877	4.783	4.707	2.036	3.717
	MI	11.433	9.975	16.230	46.693	0.998	2.123
	CE	63.578	60.018	22.271	11.673	1.247	2.654
	JE	75.025	77.652	52.442	58.366	0.998	2.123
	*SE*	6.448	6.448	6.295	5.860	2.701	4.772
	MI	10.605	8.542	24.641	31.816	0.803	1.409
	CE	33.621	30.448	18.176	14.355	3.379	5.990
	JE	45.599	45.808	45.944	46.807	0.803	1.409
	*SE*	9.112	9.112	9.354	3.742	3.915	5.676
	MI	9.761	8.453	0.000	0.000	0.791	1.032
	CE	27.217	23.315	0.000	0.000	5.682	10.940
	JE	35.367	36.004	0.000	0.000	0.828	1.088
	*SE*	11.225	11.225	11.225	0.000	2.035	5.612
	MI	10.274	9.559	0.000	0.000	0.498	2.658
	CE	32.910	21.578	0.000	0.000	0.997	5.315
	JE	45.248	40.260	0.000	0.000	0.498	2.658
	*SE*	9.112	9.112	9.354	8.790	4.526	5.450
	MI	9.761	8.453	0.000	11.225	1.579	1.701
	CE	27.217	23.315	0.000	11.225	3.804	6.541
	JE	35.367	36.004	0.000	29.933	1.653	1.766
	*SE*	6.950	6.950	6.739	6.195	2.883	5.312
	MI	8.657	6.788	20.519	25.368	0.820	1.624
	CE	27.911	24.265	9.503	13.108	3.280	6.496
	JE	37.884	37.573	32.296	38.900	0.820	1.624

a) Average entropic values are considered for Csp^2–Csp^2 bonds in aromatic systems. b) Connected subgraphs. c) Sachs subgraphs. d) MACCs fingerprints. e) Substructure fingerprints. f) Atomic hydrophobicity. g) Atomic refractivity.

additive properties like the boiling point, it is desirable that peripheral bonds (or atoms) have greater weight than the internal ones.[4,27] It could be thus anticipated that the use of the adapted weighting scheme on saturated hydrocarbons should yield better correlations for the boiling point of alkanes. A simple analysis could be performed with 18 octane isomers, using any weight X, let us say, van der Waals volume and the CS source originator algorithm.

As was expected, "swapping" the roles for the vertices, to award greater weight to the terminal vertices and lesser con-

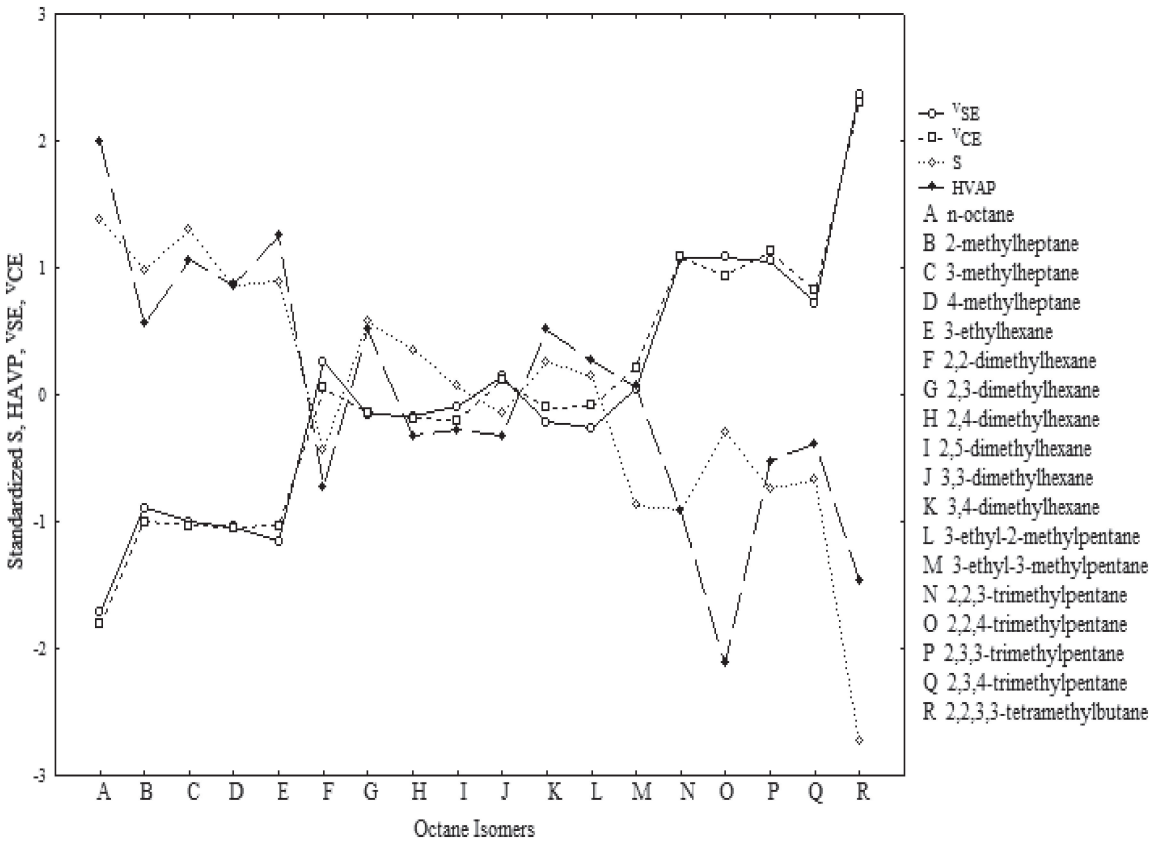


Figure 4. Tendencies for physicochemical properties and IFI values for octane isomers.

tribution to the inner ones yields improvements in univariate correlations for the boiling point of the octane isomers (for details see Table SI6, Supporting Information).

Interestingly, negative correlations are achieved with the boiling point. To comprehend this behavior, the global IFI values are analyzed. It is known that an increase in branching of alkanes results in the decrease in the degree of freedom of rotation of the C–C bonds, which lowers the molecular thermodynamics entropy. Furthermore, according to the classical thermodynamics equation for entropy and Trouton's rule,[28] entropy, boiling point and heat of vaporization are proportional terms, influenced by molecular branching (i.e., decrease with increase in ramification of the carbon skeleton).[29] Thus negative correlation should be due to a shift of the entropy values to the opposite direction with respect to these properties, which is indeed demonstrated in Figure 4. Likewise, the isomeric variation of these properties is assessed using a special graph proposed by Randić,[30] whose coordinates are paths of length two and three for octane isomers [(p2, p3) graph], revealing consistency with the trends observed in Figure 4 (see Supporting Information, SI7 and SI8).

Why moderate single variable correlations are yielded with the boiling and melting points is rather difficult to interpret. It has been suggested that it is probably due to some sort of synergetic phenomenon not interpretable by a microscopic mechanism. Also, single variable models do not yield good correlations for properties related with both the dynamic and static attributes of molecular structures like critical temperature, volume, and density. Good correlations for such properties require contributions from more than one variable. However, interpretations for such models are not straight forward. Certainly the aim of this study is not to find the best correlations, as only a very small portion of the GT-STAF descriptor space is explored. An important inference for this study is that the GT-STAF structure possesses two contrasting tendencies; one characterized by a greater role for terminal vertices and lesser one for inner vertices (unweighted entropic measures), and a reversed trend for the other (weighted entropic measures), which awards this approach notable versatility, permitting its application in the correlation studies of a wide range of molecular structural properties. Similar analyzes for other molecular "fragment" models could be performed as well. On the other hand, it has been suggested that indices that discriminate between more buried (topologically involved) vertices from terminal (exposed) ones are related with molecular accessibility areas and/or volumes and would thus codify information on intermolecular forces and interactions.[5,7,10,31]

An ideal study to evaluate this hypothesis is by means of the analysis of the amino acid conformation properties, addressed in the subsequent section. To acquire deeper insight on the information codified by the GT-STAF approach, more amino acid properties (physicochemical and thermodynamic) were analyzed as well.[32–37]

Amino Acid Properties and Their Relation to GT-STAF Indices. Several reports in the literature have pointed to the existence of a close relationship between protein stability and amino acid physicochemical, energetic and conformational properties.[32–37] Property sets of different magnitudes for the 20

coded amino acids have been proposed by various authors for use in proteomic tasks, with one of the most common being a set of 48 amino acid properties, proposed by Gromiha et al.[33]

In other cases, these properties are condensed in orthogonal "properties" using principal components analysis. Examples include: the three z-scales derived from a set of 29 physicochemical variables,[38] KOKOS descriptors (10 principal properties computed from a collection of 188 physicochemical properties),[39,40] the VHSE descriptor (comprised of 8 principal properties derived from 50 physicochemical properties of the 20 coded amino acids)[41] and 2 principal properties obtained from molecular interaction fields,[42] among others.

In this subsection, our primary aim is to explore the amino acid property space with the hope of gaining more insight on the information codified by the GT-STAF IFIs, following the synthesis that if an index correlates with a particular property, computed or experimentally determined, then that index could be interpreted in terms of that construct. To guarantee a broader amino acid property space for exploration, 70 variables we considered and these comprised of the 48 (conformational, physicochemical, and thermodynamic) properties, 3 z-scores, 8 VHSE principle properties, 2 principal properties for molecular interaction fields, 5 T-scale properties,[43] in addition to the isotropic surface area (ISA),[44] the electronic charge index (ECI),[44] and the hydropatic indices (HWS and KDS).[45,46] Likewise, 12 indices were calculated for each molecular "fragment" model (3 global invariant operators, i.e. one metric, mean and statistical invariant, on the 4 atomic entropy measures, i.e. *SE*, *MI*, *CE*, and *JE*), using the van der Waals volume (V) as weight, to yield a total of 132 GT-STAF indices.

To explore the information codified by the GT-STAF approach, factor analysis was carried out for a data set comprised of the amino acid properties and the GT-STAF indices and "varimax normalized" is used as the rotational strategy to allow easy interpretation of the factor loadings. The theoretical aspects of this statistical method have been explained elsewhere.[47–49] In this analysis, only factor loadings greater than 0.60 are considered. The factor loadings, eigenvalues and the percentages of the explained variance by ten principal factors obtained in this study are available as Supporting Information (SI9 and SI10).

Factor 1 (30.761%) possesses robust loadings for IFIs for molecular "fragment" models based on CS, Q, K, VP, S, MA, ES, and SS fingerprints, as well as ALOGP and AMR criteria, particularly computed using the summation operator as global invariant (N1). Likewise, several amino acid properties are loaded in this factor, among which are: parameters related with size and steric features [molecular weight (Mw), bulkiness (B_1), partial specific volume (V^0), refractive index (μ), helical contact area (C_a), volume (V), principal component score on steric properties ($VHSE_3$)]; hydrophobic scales [Thermodynamic transfer hydrophobicity (H_t), solvent-accessible surface area for denatured protein (ASA_D), solvent-accessible surface area for unfolding protein (ΔASA), principal property from molecular interaction fields (PP2 Hydroph.), isotropic surface area (ISA)]; thermodynamic properties [short- and medium-range nonbonded energy (E_{sm}), unfolding enthalpy change of hydration (ΔH_h), unfolding entropy change of hydration ($-T\Delta S_h$), unfolding hydration heat capacity change (ΔC_{ph}), unfolding enthalpy change of chain (ΔH_c), unfolding entropy changes of chain ($-T\Delta S_c$) and unfolding Gibbs free energy change (ΔG)] and principal components on topological indices (T1 and T5). This result indicates that there exist correlation between this set of diverse amino acid properties and the molecular "fragment" models loaded in this factor, suggesting that the latter codify information related with hydrophobic, energetic and physical features of these molecules.

Similarly, GT-STAF IFIs are strongly loaded in factor 2 (22.637%), specifically indices derived from TP and VP source originator algorithms, as well as CS and S models based on the potential mean (P3) and the kurtosis (K) operators. In the same factor is loaded the Z_3-score (contains information for amino acid properties pK_a, pI, ^{1}H NMR and some hydrophobicity/hydrophilicity scales), extracted by principal component analysis of 29 physicochemical properties for the 20 coded amino acids, suggesting that there exist a relationship between the properties important for the Z_3-score and the GT-STAF IFIs. Given that these properties are generally related with electronic and steric configurations of molecular structures, it could be deduced that the GT-STAF IFIs derived from the TP, VP, CS, and S molecular "fragment" models codify information important for these features.

Factor 3 (10.993%) is important for amino acid properties whose information is not adequately codified by single variable GT-STAF IFI correlations, particularly hydrophobicity-related parameters such as: surrounding hydrophobicity (H_p), average number of surrounding residues (N_s), combined surrounding hydrophobicity (H_{gm}), solvent-accessible surface area for native protein (ASA_N), buriedness (Br), normalized consensus hydrophobicity (H_{nc}), solvent-accessible reduction ratio (R_a) as well as the principal component on hydrophobicity ($VHSE_1$) and hydrophilicity (Z_1); conformational measures: long-range contacts (N_l), flexibility (number of side-chain dihedral angles) (f); thermodynamic properties: total nonbonded energy (E_t), long-range nonbonded energy (E_l), Gibbs free energy change of hydration for unfolding (ΔG_h), free energy change of hydration for native protein (G_{hN}), unfolding entropy change of protein ($-T\Delta S$), unfolding enthalpy change (ΔH), as well as other physicochemical properties such as: polarity (P) and related principal component (PP1), chromatographic index (R_f); electronic parameters: the electronic charge index (ECI) and principal component on electronic properties (VHSE8) and finally the principal component on topological descriptors (T2).

On the other hand, factor 5 (4.448%) reveals strong correlation for the information index ^VJE_K (ES) [JE computed on the ES molecular "fragment" model and the Kurtosis operator] and amino acid properties pH at isoelectric point (pHi) and the propensity to be at C-terminal (α_c).

Note that factors 4 (8.508%) and 6 (4.112%) are exclusive for the GT-STAF methodology, implying that there is important structural information codified by the proposed IFIs that is orthogonal to the considered amino acid properties.

Based on the evidence provided by this experiment, it could be deduced that the GT-STAF approach codifies important physical, physicochemical, conformational, and energetic features of molecules, demonstrating its versatility.

Table 4. Statistical Parameters of Best Models for Ionization Constant of Aniline Derivatives

Variables	n	R^2	Q^2	s	F
Hammett σ constants[59]	1	0.940	—	0.310	530
Natural charge (Q_n)[59]	1	0.915	—	0.368	365
Relative proton-transfer enthalpy (ΔH_{prot})[59]	1	0.921	—	0.354	398
Minimum ionization energy ($I_{S,min}$)[59]	1	0.949	—	0.285	633
Minimum molecular electrostatic potential (V_{min})[59]	1	0.945	—	0.300	585
GT-STAF (eq 9)	1	0.720	0.688	0.677	87.59
GT-STAF (eq 9)	2	0.863	0.835	0.481	103
GT-STAF (eq 10)	3	0.926	0.904	0.359	133
GT-STAF (eq 11)	4	0.949	0.932	0.302	145

Evaluation of Relationship between Experimental pK_a Values of Substituted Anilines and GT-STAF Indices. The effect of substituents on the reactivity and properties of chemical structures constitutes an area of particular interest in structure activity relationship studies. Several parameters have been successfully employed in the prediction of chemical properties due to variations in substituent groups which include empirical parameters (e.g. Hammett constant),[50–52] quantum chemical descriptors (e.g. the molecular electrostatic potential),[53–55] as well as group contribution methods (e.g. the Free–Wilson model),[56] among others. Among the extensively studied properties is the ionization constant (pK_a) of organic acids and bases.[52,57–59] Proton uptake or release controls many important biological processes. Proper pK_a predictions are of critical importance in structure activity studies given that many biological processes such as respiration, catalysis and protein synthesis entail proton exchange mechanisms.

This section aimed at assessing the possible relationship between the acid ionization constant (pK_a) and the GT-STAF approach through univariate correlations, using 36 aniline derivatives. This group of chemical compounds has been used by several authors to study the predictive ability of the ionization constant using different molecular parameters.[59,60]

As it is evident from Table 4, only moderate correlation ($R^2 = 0.720$) is obtained with univariate models. This however does not imply that good correlations are not obtainable with the GT-STAF approach, but rather indicates that all the information relevant for pK_a correlations is not condensable in single variables, and thus modeling of such a property requires contributions from more than one variable. This is indeed true, as bivariate and three variable models yield improved correlation coefficients (R^2) of 0.863, 0.926, and 0.949 for two, three and four variable models, respectively, with the latter comparable to the best correlation obtained for quantum chemical parameters,[59] see Table 4 (for GT-STAF regression model equations, see Supporting Information, SI11).

The low F value for the GT-STAF model is because this parameter is inversely proportional to the degree of freedom for the regression model. Up to this point, correlations of the different parameters with the ionization constant (pK_a) have been analyzed. To obtain a clearer picture of the existent relationship for the response variable, the quantum-chemical descriptors and the GT-STAF approach (17 variables are considered), principal component analysis is performed (the variables and the factor loadings for this study are available as Supporting Information, SI12). Two important aspects are observed: 1) the quantum chemical descriptors are strongly correlated with the studied property evidenced by robust loadings in the same factor, and thus capture the same essence of the chemical structure, and 2) the GT-STAF IFIs are mainly loaded in Factor 1 (53.75%), signifying that these indices capture some other form of structural information orthogonal to that of the pK_a and the quantum chemical parameters.

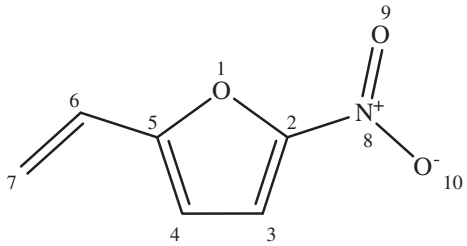
However, there are a few GT-STAF IFIs that are loaded in the same factor as quantum chemical indices. It is not surprising that it is these that yield the best correlations for pK_a (see Supporting Information, SI11). Moreover, there are loadings for GT-STAF IFIs in Factor 3 (9.51%), 4 (7.21%), and 5 (3.27%). Accordingly, the conclusion arrived at for this section is that, generally, the GT-STAF approach codifies some other form of electronic information orthogonal to empirical, quantum chemical and pK_a parameters for molecules, although contributions from various MDs provides the necessary information for adequate description of this property.

Atomic Contributions to Global Molecular Properties. In previous reports, in order to demonstrate the usability of the GT-STAF approach in QSPR studies, a search of the best models for the physicochemical property, the partition coefficient n-octanol/water, using a database of 34 derivatives of furylethylenes was performed,[12–15] demonstrating superior performance than DRAGON MDs and other models reported in the literature.[17,30]

In this case, our interest is to interpret the GT-STAF IFIs in terms of the atomic contributions to the global partition coefficient. Accordingly, a five variable model for log P (eq 8) was built using total IFIs computed using Manhattan's distance operator (equivalent to the linear combination of atomic entropy values). The analysis of atomic contributions to the global partition coefficient permits determining the atoms (or atom groups) majorly contribute to the studied property.

$$\begin{aligned}\mathbf{log}\,\boldsymbol{P} = {} & 4.242\ (1.265) - 0.960\ (\pm 0.184)\ {}^{\mathbf{V}}\mathbf{SE_{N1}[(D)]CS} + 0.091\ (\pm 0.020)\ {}^{\mathbf{V}}\mathbf{MI_{N1}[(D)]CS} \\ & - 0.013\ (\pm 0.003)\ {}^{\mathbf{V}}\mathbf{MI_{N1}[(D)]SS} + 0.249\ (\pm 0.044)\ {}^{\mathbf{V}}\mathbf{SE_{N1}[(D)]S} + 0.347\ (\pm 0.074)\ {}^{\mathbf{V}}\mathbf{SE_{N1}[(D)]V} \qquad (8) \\ R^2 = 81.63,\ & Q^2\mathrm{loo} = 74.53,\ F = 24.89,\ s = 0.33\end{aligned}$$

Table 5. Atomic Contributions to the Global Partition Coefficient of 2-Nitro-5-vinylfuran Using Equation 12



Atom	$^{V}SE_{N1}$[(D)]CS	$^{V}MI_{N1}$[(D)]CS	$^{V}MI_{N1}$[(D)]SS	$^{V}SE_{N1}$[(D)]S	$^{V}SE_{N1}$[(D)]V	log *P* (eq 12)	G and C Scale
O_1	0.280	0.583	0.779	0.248	0.276	0.356	0.391
C_2	2.324	4.840	10.660	2.075	2.039	−0.278	0.138
C_3	2.607	5.584	14.400	2.597	2.902	−0.100	−0.045
C_4	2.531	5.772	14.400	2.571	3.021	0.025	−0.045
C_5	2.195	5.195	11.621	2.109	2.139	−0.091	0.138
C_6	2.282	6.753	12.142	2.647	2.573	0.245	−0.036
C_7	2.316	8.442	17.273	3.251	2.268	0.347	−0.239
N_8	0.511	1.232	1.397	0.546	0.458	0.323	−3.185
O_9	0.183	0.556	1.146	0.174	0.165	0.385	1.824
O_{10}	0.366	1.112	0.648	0.347	0.331	0.367	1.824
Global	15.595	40.069	70.066	16.565	16.172	1.579	0.765

The molecular structure of 2-nitro-5-vinylfuran was selected as reference example for the study of atomic contributions to the global molecular partition coefficient (Table 5). Although these contributions may not represent the "actual" constructive algorithm for molecular hydrophobicity, this result offers valuable understanding of the GT-STAF approach, in terms of the weight placed on contributions of different fragments. However, these "shifts" rather than their magnitudes are to a greater extent consistent with chemical intuition. Firstly, it is known that though the nitro group is polar in nature and thus with a strong hydration effect, when attached to an aromatic ring system, it attains a positive hydrophobic contribution. The justification for this trend is the stabilizing effect due the interaction between nonbinding nitro group electrons and the π cloud of aromatic system which lowers the molecular system entropy. Likewise, the extended chain-conjugation with the furyl group lowers the hydrational effect due an aqueous medium, which results in an increase in the hydrophobicity.

On the other hand, the negative shift for the furyl group on the GT-STAF hydrophobicity scale seems somewhat "misplaced." In reality, this means that the information codified by the GT-STAF approach represents a component important for molecular hydrophobicity, characterized by a reductionist tendency for aromatic systems resulting in less lipophilic behavior. A comparison with the Ghose and Crippen atomic hydrophobicity parameters reveals a quite different scheme (Table 5), giving greater weight to the furyl group (0.577) than the nitro group (0.463), while a negative contribution is attributed to the ethylene side chain (−0.275). This reveals that the GT-STAF and the Ghose and Crippen atomic hydrophobicity scales are unrelated and thus the former codifies some other form of information equally important for computing the molecular hydrophobicity (Note that for the Ghose and Crippen scale includes hydrogen atom contributions, and thus the total value is less than the actual value when hydrogen atoms are considered). The assessment of the atom or fragment contributions to molecular properties permits the elucidation of the structure–activity relationships (SAR) of chemical entities, which in turn rationalizes the mechanistic modification of chemical structures to enhance (or reduce) the magnitude of their properties or activities, as well as the identification of bioisosteric molecular fragments.

Conclusion

In this report, several approaches aimed at understanding the information codified by the GT-STAF IFIs are presented. These studies demonstrate that while most chemical parameters, say, the boiling point or molecular mass, have single interpretations due to their chemical or physical definitions, theoretical MDs capture diverse structural, chemical, and physicochemical information and would thus be too limiting to analyze these indices from a unique model. It should also be remembered that, not all chemical information captured by mathematical algorithms is interpretable using known models and as new models are discovered, other interpretations could be given to these theoretical parameters. However, if an index depicts trends consistent with a traditional chemical model, then it could be interpreted in terms of this model. The inferences from these studies should provide plausible guidelines for the mechanistic interpretation of chemical property mathematical models built with the GT-STAF indices, as well as for the modification or design of chemical structures with atoms/fragments that provide the desired contributions to global molecular properties.

Barigye, S. J. and Freitas, M.P. acknowledge financial support from CNPq. Marrero-Ponce, Y. thanks the program "*International Professor*" for a fellowship to work at Cartagena University in 2013–2014.

Supporting Information

Tables, figures, and model equations as supplementary information material (SI1–SI12), as well as the datasets used in each study (SI13) are available free of charge on J-STAGE.

References

1 H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17.
2 M. Randić, *J. Math. Chem.* **1996**, *19*, 375.
3 I. Motoc, A. T. Balaban, *Rev. Roum. Chim.* **1981**, *26*, 593.
4 M. Randić, J. Zupan, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 550.
5 E. Estrada, *Internet Electron. J. Mol. Des.* **2002**, *1*, 360.
6 Q.-N. Hu, Y.-Z. Liang, H. Yin, X.-L. Peng, K.-T. Fang, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1193.
7 E. Estrada, *J. Phys. Chem. A* **2002**, *106*, 9085.
8 J. K. Labanowski, I. Motoc, R. A. Dammkoehler, *Comput. Chem.* **1991**, *15*, 47.
9 L. B. Kier, L. H. Hall, *J. Pharm. Sci.* **1981**, *70*, 583.
10 L. B. Kier, L. H. Hall, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 792.
11 R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, 1st ed., WILEY-VCH, Weinheim, **2009**, Vol. 2, p. 667.
12 S. J. Barigye, Y. Marrero-Ponce, O. Martinez Santiago, Y. Martinez Lopez, F. Perez-Gimenez, F. Torrens, *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 164.
13 S. J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, F. Torrens, L. M. Artiles-Martínez, R. W. Pino-Urias, O. Martínez-Santiago, *J. Comput. Chem.* **2013**, *34*, 259.
14 S. J. Barigye, Y. Marrero-Ponce, Y. M. López, O. M. Santiago, F. Torrens, R. G. Domenech, J. Galvez, *SAR QSAR Environ. Res.* **2013**, *24*, 3.
15 S. J. Barigye, Y. Marrero-Ponce, V. Alfonso-Reguera, F. Pérez-Giménez, *Chem. Phys. Lett.* **2013**, *570*, 147.
16 S. J. Barigye, Y. Marrero-Ponce, F. Pérez-Giménez, D. Bonchev, *Mol. Diversity* **2014**, *18*, 673.
17 D. Bonchev, N. Trinajstic, *J. Chem. Phys.* **1977**, *67*, 4517.
18 C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379; C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 623.
19 T. M. Cover, J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, Hoboken, New Jersey, **2006**.
20 E. Desurvire, *Classical and Quantum Information Theory*, Cambridge University Press, New York, **2009**.
21 A. T. Balaban, in *Symmetry: Unifying Human Understanding*, ed. by I. Hargittai, Pergamon Press, New York, **1986**, pp. 999–1020.
22 R. M. Dannenfelser, N. Surendran, S. H. Yalkowsky, *SAR QSAR Environ. Res.* **1993**, *1*, 273.
23 P. G. Mezey, *J. Am. Chem. Soc.* **1990**, *112*, 3791.
24 P. G. Mezey, *Int. J. Quantum Chem.* **1990**, *38*, 699.
25 P. G. Mezey, in *Concepts and Application of Molecular Similarity*, ed. by M. A. Johnson, G. M. Maggiora, Wiley-VCH Verlag GmbH, Weinheim, Germany, **1990**, pp. 321–368.
26 L. B. Kier, L. H. Hall, *Pharm. Res.* **1990**, *7*, 801.
27 M. Randić, *Croat. Chem. Acta* **2004**, *77*, 1.
28 R. Chang, *Physical Chemistry for the Biosciences*, University Science Books, Sansalito, CA, **2005**, p. 677.
29 H. Narumi, H. Hosoya, *Bull. Chem. Soc. Jpn.* **1985**, *58*, 1778.
30 M. Randić, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672.
31 J. Gálvez, *THEOCHEM* **1998**, *429*, 255.
32 K. Tomii, M. Kanehisa, *Protein Eng.* **1996**, *9*, 27.
33 M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, *J. Protein Chem.* **1999**, *18*, 565.
34 M. M. Gromiha, *J. Chem. Inf. Model.* **2005**, *45*, 494.
35 M. M. Gromiha, S. Selvaraj, A. M. Thangakani, *J. Chem. Inf. Model.* **2006**, *46*, 1503.
36 M. M. Gromiha, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1481.
37 J. Caballero, L. Fernández, J. I. Abreu, M. Fernández, *J. Chem. Inf. Model.* **2006**, *46*, 1255.
38 S. Hellberg, M. Sjöström, B. Skagerberg, S. Wold, *J. Med. Chem.* **1987**, *30*, 1126.
39 A. Kidera, Y. Konishi, M. Oka, T. Ooi, H. A. Scheraga, *J. Protein Chem.* **1985**, *4*, 23.
40 L. Pogliani, *J. Phys. Chem.* **1994**, *98*, 1494.
41 H. Mei, Z. H. Liao, Y. Zhou, S. Z. Li, *Pept. Sci.* **2005**, *80*, 775.
42 G. Cruciani, M. Baroni, E. Carosati, M. Clementi, R. Valigi, S. Clementi, *J. Chemometr.* **2004**, *18*, 146.
43 F. Tian, P. Zhou, Z. Li, *J. Mol. Struct.* **2007**, *830*, 106.
44 E. R. Collantes, W. J. Dunn, III, *J. Med. Chem.* **1995**, *38*, 2705.
45 T. P. Hopp, K. R. Woods, *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824.
46 J. Kyte, R. F. Doolittle, *J. Mol. Biol.* **1982**, *157*, 105.
47 A. Basilevsky, *Statistical Factor Analysis and Related Methods*, Wiley, New York, **1994**, p. 738.
48 E. R. Malinowski, D. G. Howery, *Factor Analysis in Chemistry*, Wiley-Interscience, New York, **1980**.
49 R. Franke, *Theoretical Drug Design Methods*, Elsevier, Amsterdam, **1984**.
50 C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178.
51 D. F. Ewing, in *Correlation Analysis in Chemistry*, ed. by N. B. Chapman, J. Shorter, Plenum Press, New York, **1978**, pp. 357–396. doi:10.1007/978-1-4615-8831-3_8.
52 A. I. Biggs, R. A. Robinson, *J. Chem. Soc.* **1961**, 388.
53 H. B. Broughton, S. M. Green, H. S. Rzepa, *J. Chem. Soc., Chem. Commun.* **1992**, 37.
54 L. Bonati, E. Fraschini, M. Lasagni, E. Palma Modoni, D. Pitea, *THEOCHEM* **1995**, *340*, 83.
55 C. W. Jefford, M. Grigorov, J. Weber, H. P. Lüthi, J. M. J. Tronchet, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 354.
56 H. Kubinyi, *J. Med. Chem.* **1976**, *19*, 587.
57 C. Mercier, O. Mekenyan, J. E. Dubois, D. Bonchev, *Eur. J. Med. Chem.* **1991**, *26*, 575.
58 T. W. Schultz, *Ecotoxicol. Environ. Saf.* **1987**, *14*, 178.
59 K. C. Gross, P. G. Seybold, Z. Peralta-Inga, J. S. Murray, P. Politzer, *J. Org. Chem.* **2001**, *66*, 6919.
60 B. G. Tehan, E. J. Lloyd, M. G. Wong, W. R. Pitt, E. Gancia, D. T. Manallack, *Quant. Struct.-Act. Relat.* **2002**, *21*, 473.

