SIMILARITY OF MOLECULAR DESCRIPTORS: THE EQUIVALENCE OF ZAGREB INDICES AND WALK COUNTS

J. Braun, A. Kerber, M. Meringer, C. Rücker[1]

Department of Mathematics,
University of Bayreuth,
D-95440 Bayreuth, Germany

Received September 20, 2004

Abstract. The similarity of 608 molecular descriptors (including topological and geometrical indices) is measured using correlation coefficients. The computations are based on a library of 10946 diverse compounds. As a result, the second Zagreb index M_2 and $mwc^{(3)}$, the molecular walk count of length 3, were found and proven to be affine dependent: $mwc^{(3)} = 2M_2$.

1. Introduction

Molecular descriptors are important tools whenever chemical compounds, usually represented by molecular graphs, have to be mapped onto real numbers. A molecular graph is a multigraph whose vertices, representing atoms, are colored by element symbols and optionally atomic state symbols, its edges represent chemical bonds and thus may be colored by bond multiplicities. Molecular descriptors are invariants of molecular graphs.

Molecular descriptors are typically applied in the analysis of quantitative structure property and structure activity relationships (QSPR/ QSAR), where they provide numerical input for supervised statistical learning. In the simplest case, molecular descriptors serve as independent variables in multiple linear regression (MLR).

Ever since new molecular descriptors were invented, their mutual interrelatedness has been an issue of practical as well as theoretical interest. Thus correlation coefficients within (mostly small) sets of topological indices have been calculated on (mostly rather homogenous) compound sets, and samples of topological indices were partitioned into clusters of similar descriptors based on diverse compound sets [1, 2, 3]. Using MOLGEN-QSPR [4] we were now able to calculate intercorrelation coefficients between many topological indices and

[1] Corresponding author e–mail: christoph.ruecker@uni-bayreuth.de

molecular descriptors of various other types (comprehensiveness in this field is unachievable) based on data for a very large and diverse set of compounds.

2. Mathematical Background

Definition 2.1. Let $\mathbf{x} = (x_1, ..., x_n)$, $\mathbf{y} = (y_1, ..., y_n) \in \mathbb{R}^n$ be vectors of values of descriptors x and y for compounds $1, ..., n$, and $\bar{x} := \frac{1}{n}\sum_i x_i$, $\bar{y} := \frac{1}{n}\sum_i y_i$ denote the arithmetic means. With $\bar{\mathbf{x}} := (\bar{x}, ..., \bar{x})$, $\bar{\mathbf{y}} := (\bar{y}, ..., \bar{y}) \in \mathbb{R}^n$ for $\mathbf{x} \neq \bar{\mathbf{x}}$ and $\mathbf{y} \neq \bar{\mathbf{y}}$ the correlation coefficient of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$r(\mathbf{x}, \mathbf{y}) := \frac{\langle \mathbf{x} - \bar{\mathbf{x}}, \mathbf{y} - \bar{\mathbf{y}} \rangle}{\|\mathbf{x} - \bar{\mathbf{x}}\| \cdot \|\mathbf{y} - \bar{\mathbf{y}}\|}.$$

Here $\langle ., . \rangle$ denotes the canonical scalar product in $\mathbb{R}^n$ defined by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$ and $\|.\|$ stands for the euclidean norm, $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. We say $\mathbf{x}$ and $\mathbf{y}$ are fully correlated if $|r(\mathbf{x}, \mathbf{y})| = 1$.

By the Cauchy–Schwarz inequation we have $|r(\mathbf{x}, \mathbf{y})| \leq 1$ and we can easily prove that $|r(\mathbf{x}, \mathbf{y})| = 1$ exactly if $\mathbf{x}$ and $\mathbf{y}$ are affine dependent, i.e. if there exist $a, b, c \in \mathbb{R}$, $a \neq 0$, $b \neq 0$ with

$$a\mathbf{x} + b\mathbf{y} + c\mathbf{1} = \mathbf{0}.$$

Furthermore, being fully correlated defines an equivalence relation on $\mathbb{R}^n$. I.e. the property of being fully correlated is reflexive, symmetric and transitive, which means that for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, $\mathbf{x} \neq \bar{\mathbf{x}}$, $\mathbf{y} \neq \bar{\mathbf{y}}$, $\mathbf{z} \neq \bar{\mathbf{z}}$ and fully correlated descriptors the following holds:

(i) reflexive: $|r(\mathbf{x}, \mathbf{x})| = 1$,
(ii) symmetric: $|r(\mathbf{x}, \mathbf{y})| = 1 \;\Rightarrow\; |r(\mathbf{y}, \mathbf{x})| = 1$ and
(iii) transitive: $|r(\mathbf{x}, \mathbf{y})| = 1 \wedge |r(\mathbf{y}, \mathbf{z})| = 1 \;\Rightarrow\; |r(\mathbf{x}, \mathbf{z})| = 1$.

Affine dependent variables contain redundant information, in particular with respect to statistical learning methods such as MLR, and can cause numerical instability. Therefore it is useful to know which pairs of variables are affine dependent, and to apply only one variable out of each equivalence class of fully correlated variables in order to keep the computational costs for the learning algorithm low.

3. Experimental

3.1. **Molecular Library.** The chemical structures for our investigation were taken from the freely available file *MayDec02CCeu.sdf*[2]. It contains 13410 diverse chemical compounds stored in *MDL SDfile* format[3]. Since many molecular descriptors cannot deal with disconnected molecular graphs, we removed such structures, as well as ions, radicals and structures containing formally charged atoms.

[2] www.maybridge.com

[3] www.mdl.com

	Minimum	1st Qu.	Median	Mean	3rd Qu.	Maximum
MW (incl. H)	67.09	186.25	242.12	257.09	312.75	1297.80
A (incl. H)	5.00	20.00	26.00	27.63	33.00	208.00
B (incl. H)	4.00	20.00	27.00	28.36	34.00	216.00
C	0.00	1.00	2.00	1.73	2.00	9.00
rings	0.00	1.00	2.00	2.07	3.00	264.00

Table 1. Characteristics of the molecular library

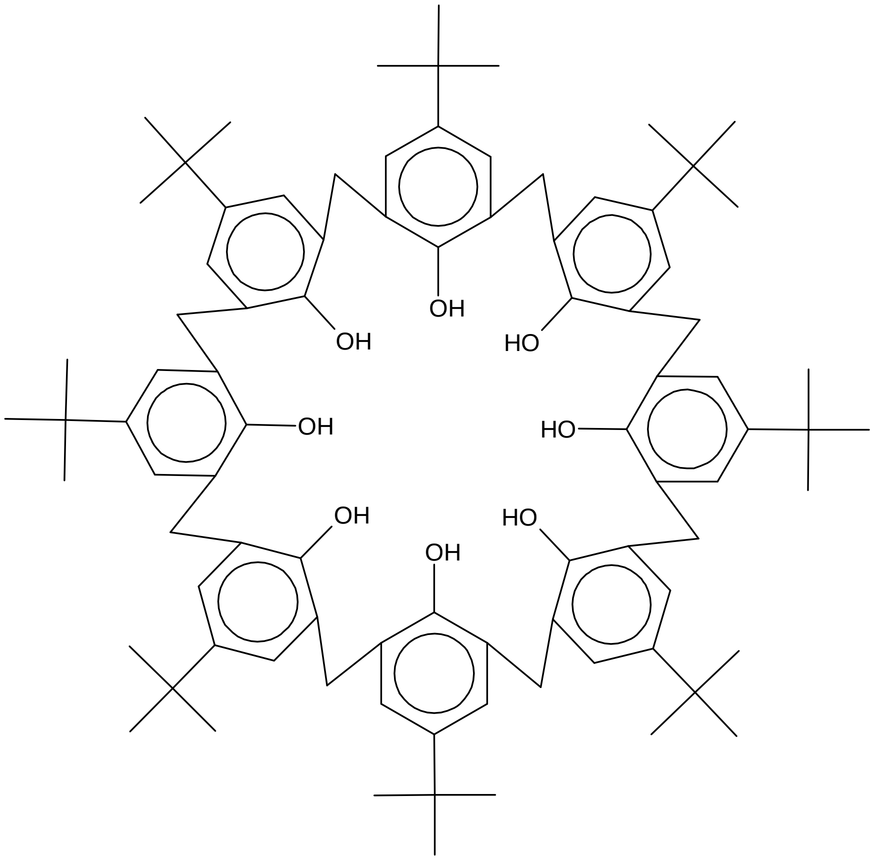


Figure 1. 4–tert–butylcalix[8]arene

We restricted our examination to structures containing only elements that are typical for organic chemistry: H, C, N, O, F, Si, P, S, Cl, Br and I. The original library also includes compounds containing B (66 structures), Na (33), K (7), Cr (1), Fe (5), Zn (2), Se (3), Ag (1), Sn (10), Te (2). These structures were excluded.

The remaining library contains 10946 compounds. Using a canonical labeling algorithm [5] we convinced ourselves that there are no duplicate structures.

Atoms	H	C	N	O	F	Si	P	S	Cl	Br	I
1	26	2	3153	3154	505	14	361	2518	2325	851	139
2	92	20	2314	3078	182	4	16	532	1291	185	16
3	222	71	1106	1554	1057	0	2	55	514	46	1
4	415	215	700	827	58	1	0	8	188	12	1
5	538	375	249	196	24	0	0	1	26	3	0
6–10	4442	4622	61	125	241	0	0	1	24	4	0
11–15	3114	3757	0	4	31	0	0	0	0	0	0
16–20	1406	1353	0	2	6	0	0	0	0	0	0
21–25	438	390	0	0	3	0	0	0	0	0	0
26–30	140	95	0	0	1	0	0	0	0	0	0
> 30	86	42	0	0	0	0	0	0	0	0	0
Sum	10919	10942	7583	8940	2108	19	379	3115	4368	1101	157

Table 2. Atom profile for the molecular library

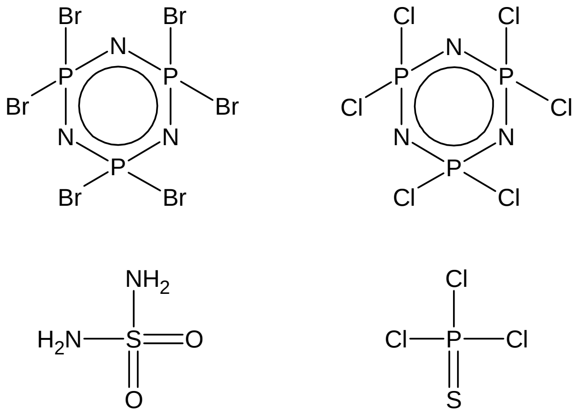


Figure 2. Compounds without carbon

The library contains 553 acyclic, 4145 carbocyclic and 6248 heterocyclic compounds (nonaromatic and aromatic) of all major functional classes: Hydrocarbons; halides, alcohols, ethers, phenols; aldehydes, ketones, acetals; carboxylic acids, carboxylic acid chlorides and anhydrides, esters, amides, nitriles; carbonates, ureas and thioureas; amines, hydrazines, hydroxylamines, nitroso compounds; thiols, thioethers, disulfides, thioacetals, sulfones, sulfonic acids, sulfonic acid chlorides, sulfonates, sulfonamides; phosphanes, phosphanoxides, phosphonates; etc.

Table 1 offers a brief description of the molecular library. Minimum, arithmetic mean, median and first/third quartiles are given for the molecular weight of the compounds, the number of atoms and bonds, the cyclomatic number and the number of rings. The compound with highest values of all these five descriptors is 4–tert–butylcalix[8]arene shown in Figure 1.

Rings	^3rings	^4rings	^5rings	^6rings	^7rings	^8rings
1	262	66	3471	5386	132	102
2	14	1	529	2724	10	0
3	0	0	34	683	5	67
4	0	0	6	198	0	0
5	0	0	0	43	0	3
6	0	0	6	13	0	1
7	0	0	0	6	2	0
8	0	0	0	3	2	2
9	0	0	0	0	0	5
≥ 10	0	0	0	0	0	1
Sum	276	67	4046	9056	151	181

Table 3. Cycle profile for the molecular library

Table 2 shows an atom profile of these compounds. The entries represent the number of compounds containing a specified number of atoms of a specified element. For instance, the library contains 4 compounds having 11–15 O atoms. Four compounds lack any carbon, phosphazenes and derivatives of sulfuric and phosphoric acid (Figure 2). The entries of Table 3 represent the number of compounds containing a specified number of rings of specified size

As usual for *SDfiles*, structures in *MayDec02CCeu.sdf* are coded without H atoms, aromatic bonds or 3D coordinates. Since many descriptors require this information, we used the tools contained in our software MOLGEN-QSPR [4] for addition of explicit H atoms, for identification of aromatic bonds, and for calculation of low-energy conformations. Details for the first two steps can be found in [6]. 9587 of the 10946 contain aromatic bonds. Calculation of 3D coordinates uses a force field similar to Allinger's MM2 [7]. For each compound the lowest in energy out of ten generated conformations was kept. Afterwards structures with high steric energy values were examined manually, and bad layouts were recalculated.

3.2. **Molecular Descriptors.** Starting point of our research are 703 molecular descriptors of various types:

- arithmetical descriptors (AD, using information coded in the compound's molecular formula),
- topological descriptors (TD, using information coded in the compound's constitution),
- geometrical descriptors (GD, using 3D information coded in the compound's configuration and conformation),
- electrotopological and AI indices (EI) [8, 9],
- overall indices (OI) [10, 11],
- Crippen indices (CI) [12].

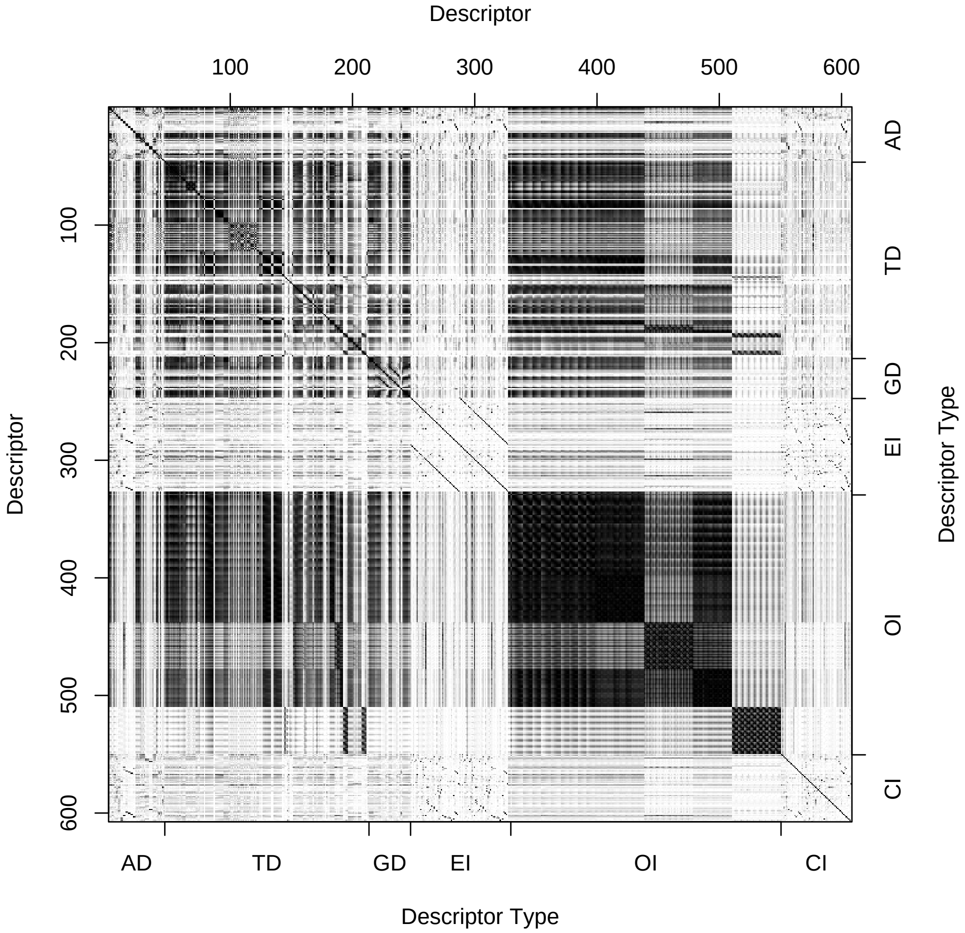


FIGURE 3. Greyscale image of the correlation matrix

Descriptor values were computed by MOLGEN–QSPR. 95 Descriptors (mainly atom counts and electrotopological state and AI indices of rare elements or atom types) turned out to be constant on the molecular library. The remaining descriptors are listed in the Appendix. The descriptors' abbreviations refer to [13], the exact definitions can be found in [14]. While descriptor calculation for the 10946 compound library took several hours, computation of the 608 × 608 correlation matrix is carried out in less than 12 min on a Windows XP system with Pentium IV 2.6 GHz CPU.

4. Results

4.1. **Computational Results.** Figure 3 shows a greyscale image of the correlation matrix. Each pixel represents a correlation coefficient $r(\mathbf{x},\mathbf{y})$, colored in black if $|r(\mathbf{x},\mathbf{y})| = 1$, or in white if $r(\mathbf{x},\mathbf{y}) = 0$. For $0 < |r(\mathbf{x},\mathbf{y})| < 1$ greyscales are used. Of course the dimension of the matrix is too high for a detailed manual analysis of the correlation

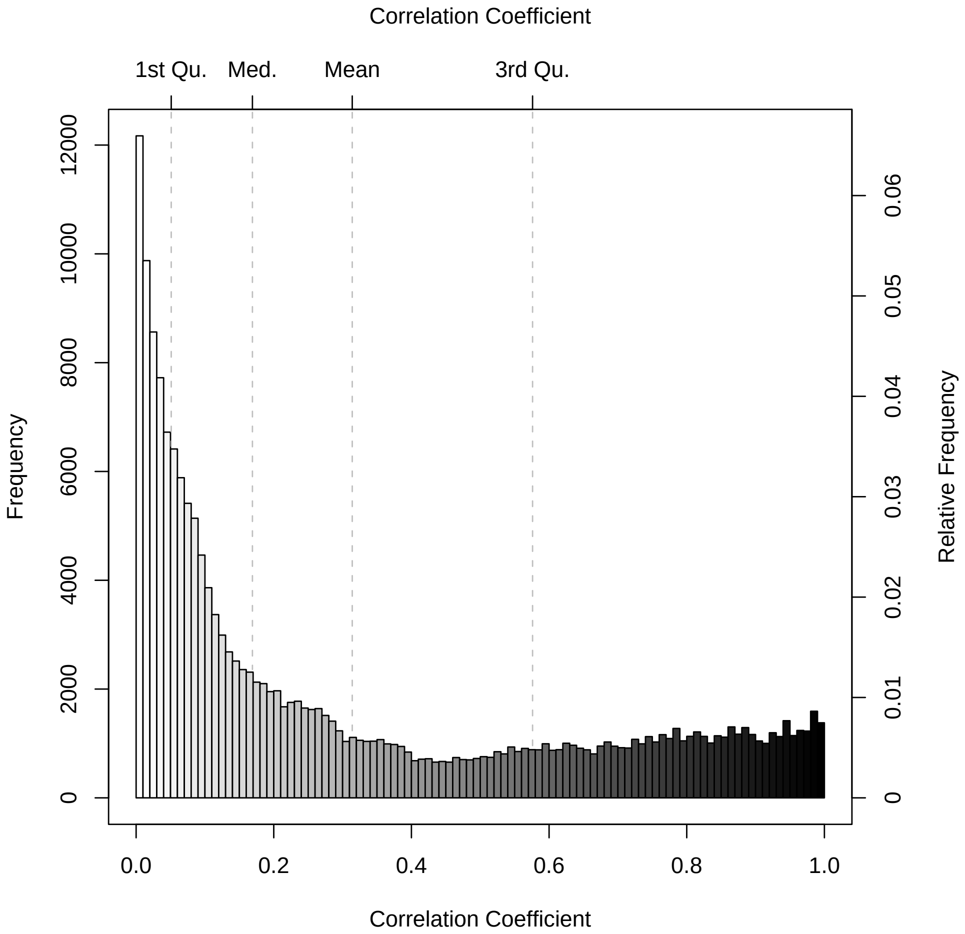


FIGURE 4. Histogram of correlation coefficients

coefficients. However, the symmetric structure of the matrix and the values 1 on the main diagonal are easily seen. For better orientation regions of the various descriptor types are indicated on the lower and right axes.

Before investigating the full correlations, let us have a closer look at the distribution of correlation coefficients. Figure 4 shows a histogram of absolute values of the correlation coefficients. Cells are filled with the same greyscales as used in Figure 3. Pairs of correlation coefficients are counted only once, diagonal elements of the correlation matrix are excluded, i.e. a total of $(608^2 - 608)/2 = 184528$ values is considered.

We have an arithmetic mean of 0.3140, the median is 0.1692, first and third quartiles are 0.05164 and 0.5764. The minimal absolute correlation coefficient is between *rel.* N_{Cl} and TM_{1c} (0.000001033).

Generally, high correlations are common among the topological indices, and in particular among the overall indices (which in fact are a special class of topological indices) and across these groups. With respect to intercorrelations, the geometrical indices as a class are not very different from the topological ones. On the other hand, electrotopological and AI indices, as well as Crippen indices, are more or less independent among themselves and of others. An exception is the high correlation between an electrotopological state index and the corresponding AI index (dark lines parallel to the main diagonal near the picture's center).

Looking for full correlations we found 26 equivalence classes of fully correlated descriptors, i.e. $|r| = 1$ with respect to double precision floating point arithmetics:

$$\begin{gathered}
A \simeq {}^0K, \\
N_F \simeq at\ F, \\
N_{Cl} \simeq at\ Cl, \\
N_{Br} \simeq at\ Br, \\
N_I \simeq at\ I, \\
N_P \simeq at\ P, \\
B \simeq {}^1K \simeq {}^0TC \simeq {}^1TC* \simeq {}^1TM_1* \simeq {}^1TM_2* \simeq {}^1TW, \\
M_1 \simeq mwc^{(2)} \simeq {}^1TC \simeq {}^0TM_1, \\
M_2 \simeq mwc^{(3)} \simeq {}^1TM_2, \\
{}^3\chi^v \simeq {}^3\chi^v_p, \\
F \simeq N_{GS} \simeq {}^2P_{acyc} \simeq {}^2P \simeq {}^2K \simeq {}^2TC* \simeq {}^2TM_1* \simeq {}^2TM_2* \simeq {}^2TW, \\
twc \sim twc_{unsat}, \\
{}^3P_{acyc} \simeq {}^3TC*_p \simeq {}^3TM_1*_p \simeq {}^3TM_2*_p \simeq {}^3TW_p, \\
{}^4P_{acyc} \simeq {}^4TC*_p \simeq {}^4TM_1*_p \simeq {}^4TM_2*_p \simeq {}^4TW_p, \\
{}^5P_{acyc} \simeq {}^5TC*_p \simeq {}^5TM_1*_p \simeq {}^5TM_2*_p \simeq {}^5TW_p, \\
{}^6P_{acyc} \simeq {}^6TC*_p \simeq {}^6TM_1*_p \simeq {}^6TM_2*_p \simeq {}^6TW_p, \\
{}^3rings \simeq {}^3TC*_{ch} \simeq {}^3TM_1*_{ch} \simeq {}^3TM_2*_{ch} \simeq {}^3TW_{ch}, \\
{}^3K \simeq {}^3TC*, \\
{}^4K \simeq {}^4TC*, \\
{}^5K \simeq {}^5TC*, \\
{}^6K \simeq {}^6TC*, \\
{}^3TC*_c \simeq {}^3TM_1*_c \simeq {}^3TM_2*_c \simeq {}^3TW_c, \\
{}^4TC*_c \simeq {}^4TM_1*_c \simeq {}^4TM_2*_c \simeq {}^4TW_c, \\
{}^6TC*_c \sim {}^6TM_1*_c \sim {}^6TM_2*_c \sim {}^6TW_c, \\
{}^4TC*_{pc} \simeq {}^4TM_1*_{pc} \simeq {}^4TM_2*_{pc} \simeq {}^4TW_{pc}, \\
{}^4TC*_{ch} \simeq {}^4TW_{ch}.
\end{gathered}$$

However, due caution is advisable in interpreting these $|r| = 1$ correlations. Thus, numerical problems in the calculations, rounding errors, as well as peculiarities of the particular compound library considered may deceptively lead to $|r| = 1$. We therefore checked all these correlations by recourse to their respective descriptor definitions. As a result, most of the above full correlations were found to be valid in general. These are indicated by the '$\simeq$' symbol. For $M_2 \simeq mwc^{(3)}$ see below.

On the other hand, a few of those listed above are certainly not full correlations in general. Thus, twc and twc_{unsat} are not at all fully correlated in general, as is easily seen for a small library of a saturated hydrocarbon and its unsaturated analogs. The full correlation of these two descriptors found in our library is due to the very broad variation of molecular size (and therefore of twc) together with comparatively small differences between twc and twc_{unsat} for some unsaturated compounds, and presumably rounding errors. At the other extreme, for isomers (constant size molecules) the same descriptors may weakly correlate; for the 217 structural isomers of benzene, $\mathsf{C_6H_6}$, twc and twc_{unsat} show a mere $r = 0.13521$.

As another example, the full correlations between the overall descriptors for 6–edge cluster subgraphs (third last line in the above list) are easily explained by the lack of compounds in our library that contain 6–edge star subgraphs, e.g. $\mathsf{SF_6}$ etc. Such correlations are indicated above by the '$\sim$' symbol.

Conversely, we cannot exclude the possibility that numerical problems may have obscured a true $|r| = 1$. Therefore we list here pairs of very highly correlated descriptors ($|r| > 0.99999$):

$$\begin{array}{rl}
P_{acyc} \sim P & (0.999997700), \\
{}^{\geq 9}P_{acyc} \sim {}^{\geq 9}P & (0.999998317), \\
{}^{4}TM_1{*} \sim {}^{4}TM_2{*} & (0.999992640), \\
{}^{5}TC{*}_c \sim {}^{5}TM_1{*}_c & (0.999996931), \\
{}^{5}TC{*}_c \sim {}^{5}TM_2{*}_c & (0.999999158), \\
{}^{5}TC{*}_c \sim {}^{5}TW_c & (0.999999558), \\
{}^{5}TM_1{*}_c \sim {}^{5}TM_2{*}_c & (0.999999304), \\
{}^{5}TM_1{*}_c \sim {}^{5}TW_c & (0.999994160), \\
{}^{5}TM_2{*}_c \sim {}^{5}TW_c & (0.999997496), \\
{}^{5}TM_1{*}_{pc} \sim {}^{5}TM_2{*}_{pc} & (0.999992664).
\end{array}$$

Correlation coefficients are given in parentheses. Note that highly correlated descriptors do not form equivalence classes, as high correlations are not transitive in general. By the descriptor definitions, none of these is a full correlation in general. The high correlations found for our compound library between the overall descriptors for 5–edge cluster subgraphs are due to few occurrences of 5–edge star subgraphs.

4.2. **Mathematical Results.** Most of the full correlations found are quite trivial. However the full correlation of the second Zagreb index M_2 and molecular walk count $mwc^{(3)}$ of length 3 seems to be formerly unknown, as it was not reported in recent surveys on Zagreb indices [3, 15].

For two descriptors to be affine dependent, full correlation on a large library is only a necessary condition. In order to make a general statement one still has to prove the affine dependence mathematically. Therefore we need the definitions of M_2 [16] and $mwc^{(3)}$ [17, 18]. As both indices do not consider chemical elements or bond multiplicities occurring in molecular graphs, they can also be regarded as invariants of simple graphs.

Definition 4.1. Let $A = (a_{ij})$ denote the adjacency matrix of a simple graph, $A^p = (a_{ij}^{(p)})$ the p-th power of the adjacency matrix, d_i the degree of vertex i and E the set of edges. Then the second Zagreb Index is defined as

$$M_2 = \sum_{\{i,j\}\in E} d_i d_j$$

and the molecular walk count of length p as

$$mwc^{(p)} = \sum_{i,j} a_{ij}^{(p)}.$$

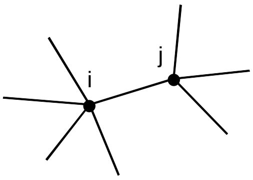
Proposition 4.2. $mwc^{(3)} = 2M_2$.

Proof.

$$\begin{aligned} mwc^{(3)} &= \sum_{i,j} a_{ij}^{(3)} = \sum_{i,j}\sum_{k} a_{ik}a_{kj}^{(2)} \\ &= \sum_{i,j}\sum_{k} a_{ik}\sum_{l} a_{kl}a_{lj} = \sum_{i,j,k,l} a_{ik}a_{kl}a_{lj} \\ &= \sum_{k,l,i,j} a_{kl}a_{ik}a_{lj} = \sum_{k,l} a_{kl}\left(\sum_{i} a_{ik}\right)\left(\sum_{j} a_{lj}\right) \\ &= \sum_{k,l} a_{kl}d_k d_l = 2\sum_{\{k,l\}} a_{kl}d_k d_l = 2\sum_{\{k,l\}\in E} d_k d_l = 2M_2 \end{aligned}$$

□

An alternative and visual proof of Proposition 4.2 follows:



Adjacent atoms i and j in this drawing (part of a larger graph) have degrees d_i and d_j, respectively. The number of walks of length 3 extending from any neighbor of atom i to atom i, then to atom j, then to any neighbor of atom j, is $d_i \cdot 1 \cdot d_j$. The number of walks of length 3 extending conversely from a neighbor of atom j to atom j, then to atom i, then to a neighbor of atom i, is $d_j \cdot 1 \cdot d_i$. The sum of these, $2d_i d_j$, is the number of all walks of length 3 that walk along bond $\{i, j\}$ in their second step. We attribute these walks to bond $\{i, j\}$. Every walk of length 3 is in this manner attributed to a bond.

Thus the number of all walks of length 3 in the molecule is the sum of these contributions over all bonds

$$mwc^{(3)} = \sum_{\{i,j\}} 2d_i d_j = 2 \sum_{\{i,j\}} d_i d_j = 2M_2$$

This elementary derivation is analogous to that of equation $mwc^{(2)} = M_1$ given in [19]. Thus both Zagreb indices may be considered special cases of the concept of walks in molecular graphs.

5. Appendix

Arithmetical descriptors: A, *A (incl. H)*, N_H, *rel.* N_H, N_C, *rel.* N_C, N_O, *rel.* N_O, N_N, *rel.* N_N, N_S, *rel.* N_S, N_F, *rel.* N_F, N_{Cl}, *rel.* N_{Cl}, N_{Br}, *rel.* N_{Br}, N_I, *rel.* N_I, N_P, *rel.* N_P, B, *B (incl. H)*, *loc. B*, *loc. B (incl. H)*, *n−*, *rel. n−*, *n− (incl. H)*, *rel. n− (incl. H)*, *n=*, *rel. n=*, *rel. n= (incl. H)*, *n#*, *rel. n#*, *rel. n# (incl. H)*, *n aromatic*, *rel. n aromatic*, *rel. n aromatic (incl. H)*, C, MW, *mean* AW, MW *(incl. H)*, *mean* AW *(incl. H)*, HBD, HBA (46 descriptors).

Topological descriptors: W, M_1, M_2, ${}^mM_1$, ${}^mM_2$, ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^0\chi^s$, ${}^1\chi^s$, ${}^2\chi^s$, ${}^3\chi^s$, ${}^3\chi^s$ *(cluster)*, ${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$, ${}^3\chi^v$, ${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$, $\Phi_{\bar{\alpha}}$, ${}^1\kappa_\alpha$, ${}^2\kappa_\alpha$, ${}^3\kappa_\alpha$, Φ, F, N_{GS}, J, J_{unsat}, MTI, MTI', H, twc, $mwc^{(2)}$, $mwc^{(3)}$, $mwc^{(4)}$, $mwc^{(5)}$, $mwc^{(6)}$, $mwc^{(7)}$, $mwc^{(8)}$, twc_{unsat}, $mwc^{(2)}_{unsat}$, $mwc^{(3)}_{unsat}$, $mwc^{(4)}_{unsat}$, $mwc^{(5)}_{unsat}$, $mwc^{(6)}_{unsat}$, $mwc^{(7)}_{unsat}$, $mwc^{(8)}_{unsat}$, G_1 *(topo. dist.)*, G_1 *(topo. dist., incl. H)*, G_2 *(topo. dist.)*, G_2 *(topo. dist.,incl. H)*, IC_0, TIC_0, CIC_0, $N*CIC_0$, SIC_0, $N*SIC_0$, BIC_0, $N*BIC_0$, IC_1, TIC_1, CIC_1, $N*CIC_1$, SIC_1, $N*SIC_1$, BIC_1, $N*BIC_1$, IC_2, TIC_2, CIC_2, $N*CIC_2$, SIC_2, $N*SIC_2$, BIC_2, $N*BIC_2$, MSD, w, w_{diag}, P_{acyc}, ${}^2P_{acyc}$, ${}^3P_{acyc}$, ${}^4P_{acyc}$, ${}^5P_{acyc}$, ${}^6P_{acyc}$, ${}^7P_{acyc}$, ${}^8P_{acyc}$, ${}^{\geq 9}P_{acyc}$, P, ${}^2P$, ${}^3P$, ${}^4P$, ${}^5P$, ${}^6P$, ${}^7P$, ${}^8P$, ${}^{\geq 9}P$, *rings*, ${}^3rings$, ${}^4rings$, ${}^5rings$, ${}^6rings$, ${}^7rings$, ${}^8rings$, ${}^{\geq 9}rings$, *ch.* G_1, *ch.* G_2, *ch.* G_3, *ch.* G_4, *ch.* G_5, *ch.* G_6, *ch.* G_7, *ch.* G_8, *ch.* J_1, *ch.* J_2, *ch.* J_3, *ch.* J_4, *ch.* J_5, *ch.* J_6, *ch.* J_7, *ch.* J_8, *ch.* J, D, ξ^c, λ_1^A, $SCA1$, $SCA2$, $SCA3$, λ_1^D, χ_T, T_m, T_3, FRB, SZD, $SZDp$, ${}^3\chi_p$, ${}^4\chi_p$, ${}^5\chi_p$, ${}^6\chi_p$, ${}^3\chi_c$, ${}^4\chi_c$, ${}^5\chi_c$, ${}^6\chi_c$, ${}^4\chi_{pc}$, ${}^5\chi_{pc}$, ${}^6\chi_{pc}$, ${}^3\chi_{ch}$, ${}^4\chi_{ch}$, ${}^5\chi_{ch}$, ${}^6\chi_{ch}$, ${}^3\chi^v_p$, ${}^4\chi^v_p$, ${}^5\chi^v_p$, ${}^6\chi^v_p$, ${}^3\chi^v_c$, ${}^4\chi^v_c$, ${}^5\chi^v_c$, ${}^6\chi^v_c$, ${}^4\chi^v_{pc}$, ${}^5\chi^v_{pc}$, ${}^6\chi^v_{pc}$, ${}^3\chi^v_{ch}$, ${}^4\chi^v_{ch}$, ${}^5\chi^v_{ch}$, ${}^6\chi^v_{ch}$, sym_{top}, R (167 descriptors).

Geometrical descriptors: G_1, G_1 *(incl. H)*, G_2, G_2 *(incl. H)*, I_A, I_B, I_C, $SHDW1$, $SHDW2$, $SHDW3$, $SHDW4$, $SHDW5$, $SHDW6$, $SHDW1/SHDW2$, $SHDW1/SHDW3$, $SHDW2/SHDW3$, $ssSHDW1$, $ssSHDW2$, $ssSHDW3$, $ssSHDW4$, $ssSHDW5$, $ssSHDW6$, $ssSHDW1/SHDW2$, $ssSHDW1/SHDW3$, $ssSHDW2/SHDW3$, V_{vdw}, ρ_{vdw}, V^s_{vdw}, V_{cub}, S_{vdw}, SAS_{H_2O}, SAS_H, D_{3D}, V_{sphere} (34 descriptors).

Electrotopological and AI indices: *S(sCH3)*, *S(dCH2)*, *S(ssCH2)*, *S(tCH)*, *S(dsCH)*, *S(aaCH)*, *S(sssCH)*, *S(ddC)*, *S(tsC)*, *S(dssC)*, *S(aasC)*, *S(aaaC)*, *S(ssssC)*, *S(sNH2)*, *S(ssNH)*, *S(aaNH)*, *S(tN)*, *S(dsN)*, *S(aaN)*, *S(sssN)*, *S(ddsN)*, *S(aasN)*, *S(sOH)*, *S(dO)*, *S(ssO)*, *S(aaO)*, *S(sF)*, *S(sssP)*, *S(dsssP)*, *S(sssssP)*, *S(sSH)*, *S(dS)*, *S(ssS)*, *S(aaS)*, *S(dssS)*, *S(ddssS)*, *S(sCl)*, *S(sBr)*, *S(sI)*, *S(ssssSi)*, *AI(sCH3)*, *AI(dCH2)*, *AI(ssCH2)*, *AI(tCH)*, *AI(dsCH)*, *AI(aaCH)*, *AI(sssCH)*, *AI(ddC)*, *AI(tsC)*, *AI(dssC)*, *AI(aasC)*, *AI(aaaC)*, *AI(ssssC)*, *AI(sNH2)*, *AI(ssNH)*, *AI(aaNH)*, *AI(tN)*, *AI(dsN)*, *AI(aaN)*, *AI(sssN)*, *AI(ddsN)*, *AI(aasN)*, *AI(sOH)*, *AI(dO)*, *AI(ssO)*, *AI(aaO)*, *AI(sF)*, *AI(sssP)*, *AI(dsssP)*, *AI(sssssP)*, *AI(sSH)*, *AI(dS)*, *AI(ssS)*, *AI(aaS)*, *AI(dssS)*, *AI(ddssS)*, *AI(sCl)*, *AI(sBr)*, *AI(sI)*, *AI(ssssSi)*, Xu, Xu^m (82 descriptors).

Overall indices: ${}^{0-8}K$, ${}^0K$, ${}^1K$, ${}^2K$, ${}^3K$, ${}^4K$, ${}^5K$, ${}^6K$, ${}^7K$, ${}^8K$, ${}^0TC$, ${}^1TC$, ${}^2TC$, ${}^3TC$, ${}^4TC$, ${}^5TC$, ${}^6TC$, TC, ${}^1TC*$, ${}^2TC*$, ${}^3TC*$, ${}^4TC*$, ${}^5TC*$, ${}^6TC*$, $TC*$, ${}^0TC^v$, ${}^1TC^v$, ${}^2TC^v$,

$^{3}TC^{v}$, $^{4}TC^{v}$, $^{5}TC^{v}$, $^{6}TC^{v}$, TC^{v}, $^{0}TM_1$, $^{1}TM_1$, $^{2}TM_1$, $^{3}TM_1$, $^{4}TM_1$, $^{5}TM_1$, $^{6}TM_1$, TM_1, $^{1}TM_1*$, $^{2}TM_1*$, $^{3}TM_1*$, $^{4}TM_1*$, $^{5}TM_1*$, $^{6}TM_1*$, TM_1*, $^{1}TM_2$, $^{2}TM_2$, $^{3}TM_2$, $^{4}TM_2$, $^{5}TM_2$, $^{6}TM_2$, TM_2, $^{1}TM_2*$, $^{2}TM_2*$, $^{3}TM_2*$, $^{4}TM_2*$, $^{5}TM_2*$, $^{6}TM_2*$, TM_2*, ^{1}TW, ^{2}TW, ^{3}TW, ^{4}TW, ^{5}TW, ^{6}TW, TW, $^{3}TC_p$, $^{4}TC_p$, $^{5}TC_p$, $^{6}TC_p$, TC_p, $^{3}TC*_p$, $^{4}TC*_p$, $^{5}TC*_p$, $^{6}TC*_p$, $TC*_p$, $^{3}TC^{v}_{p}$, $^{4}TC^{v}_{p}$, $^{5}TC^{v}_{p}$, $^{6}TC^{v}_{p}$, TC^{v}_{p}, $^{3}TM_{1p}$, $^{4}TM_{1p}$, $^{5}TM_{1p}$, $^{6}TM_{1p}$, TM_{1p}, $^{3}TM_1*_p$, $^{4}TM_1*_p$, $^{5}TM_1*_p$, $^{6}TM_1*_p$, TM_1*_p, $^{3}TM_{2p}$, $^{4}TM_{2p}$, $^{5}TM_{2p}$, $^{6}TM_{2p}$, TM_{2p}, $^{3}TM_2*_p$, $^{4}TM_2*_p$, $^{5}TM_2*_p$, $^{6}TM_2*_p$, TM_2*_p, $^{3}TW_p$, $^{4}TW_p$, $^{5}TW_p$, $^{6}TW_p$, TW_p, $^{3}TC_c$, $^{4}TC_c$, $^{5}TC_c$, $^{6}TC_c$, TC_c, $^{3}TC*_c$, $^{4}TC*_c$, $^{5}TC*_c$, $^{6}TC*_c$, $TC*_c$, $^{3}TC^{v}_{c}$, $^{4}TC^{v}_{c}$, $^{5}TC^{v}_{c}$, $^{6}TC^{v}_{c}$, TC^{v}_{c}, $^{3}TM_{1c}$, $^{4}TM_{1c}$, $^{5}TM_{1c}$, $^{6}TM_{1c}$, TM_{1c}, $^{3}TM_1*_c$, $^{4}TM_1*_c$, $^{5}TM_1*_c$, $^{6}TM_1*_c$, TM_1*_c, $^{3}TM_{2c}$, $^{4}TM_{2c}$, $^{5}TM_{2c}$, $^{6}TM_{2c}$, TM_{2c}, $^{3}TM_2*_c$, $^{4}TM_2*_c$, $^{5}TM_2*_c$, $^{6}TM_2*_c$, TM_2*_c, $^{3}TW_c$, $^{4}TW_c$, $^{5}TW_c$, $^{6}TW_c$, TW_c, $^{4}TC_{pc}$, $^{5}TC_{pc}$, $^{6}TC_{pc}$, TC_{pc}, $^{4}TC*_{pc}$, $^{5}TC*_{pc}$, $^{6}TC*_{pc}$, $TC*_{pc}$, $^{4}TC^{v}_{pc}$, $^{5}TC^{v}_{pc}$, $^{6}TC^{v}_{pc}$, TC^{v}_{pc}, $^{4}TM_{1pc}$, $^{5}TM_{1pc}$, $^{6}TM_{1pc}$, TM_{1pc}, $^{4}TM_1*_{pc}$, $^{5}TM_1*_{pc}$, $^{6}TM_1*_{pc}$, TM_1*_{pc}, $^{4}TM_{2pc}$, $^{5}TM_{2pc}$, $^{6}TM_{2pc}$, TM_{2pc}, $^{4}TM_2*_{pc}$, $^{5}TM_2*_{pc}$, $^{6}TM_2*_{pc}$, TM_2*_{pc}, $^{4}TW_{pc}$, $^{5}TW_{pc}$, $^{6}TW_{pc}$, TW_{pc}, $^{3}TC_{ch}$, $^{4}TC_{ch}$, $^{5}TC_{ch}$, $^{6}TC_{ch}$, TC_{ch}, $^{3}TC*_{ch}$, $^{4}TC*_{ch}$, $^{5}TC*_{ch}$, $^{6}TC*_{ch}$, $TC*_{ch}$, $^{3}TC^{v}_{ch}$, $^{4}TC^{v}_{ch}$, $^{5}TC^{v}_{ch}$, $^{6}TC^{v}_{ch}$, TC^{v}_{ch}, $^{3}TM_{1ch}$, $^{4}TM_{1ch}$, $^{5}TM_{1ch}$, $^{6}TM_{1ch}$, TM_{1ch}, $^{3}TM_1*_{ch}$, $^{4}TM_1*_{ch}$, $^{5}TM_1*_{ch}$, $^{6}TM_1*_{ch}$, TM_1*_{ch}, $^{3}TM_{2ch}$, $^{4}TM_{2ch}$, $^{5}TM_{2ch}$, $^{6}TM_{2ch}$, TM_{2ch}, $^{3}TM_2*_{ch}$, $^{4}TM_2*_{ch}$, $^{5}TM_2*_{ch}$, $^{6}TM_2*_{ch}$, TM_2*_{ch}, $^{3}TW_{ch}$, $^{4}TW_{ch}$, $^{5}TW_{ch}$, $^{6}TW_{ch}$, TW_{ch} (221 descriptors).

Crippen indices: $at\ C01$, $at\ C02$, $at\ C03$, $at\ C04$, $at\ C05$, $at\ C06$, $at\ C07$, $at\ C08$, $at\ C09$, $at\ C10$, $at\ C11$, $at\ C12$, $at\ C13$, $at\ C14$, $at\ C15$, $at\ C16$, $at\ C17$, $at\ C18$, $at\ C19$, $at\ C20$, $at\ C21$, $at\ C22$, $at\ C23$, $at\ C24$, $at\ C26$, $at\ C27$, $at\ H01$, $at\ H02$, $at\ H03$, $at\ H04$, $at\ O01$, $at\ O02$, $at\ O03$, $at\ O04$, $at\ O05$, $at\ O06$, $at\ O09$, $at\ O10$, $at\ O11$, $at\ N01$, $at\ N02$, $at\ N03$, $at\ N04$, $at\ N05$, $at\ N06$, $at\ N07$, $at\ N08$, $at\ N09$, $at\ N11$, $at\ Cl$, $at\ Br$, $at\ I$, $at\ F$, $at\ P$, $at\ S01$, $at\ S02$, $at\ S03$, $at\ Me01$ (58 descriptors).

References

[1] S. C. Basak, B. D. Gute, and A. T. Balaban. *Interrelationship of Major Topological Indices Evidenced by Clustering.* Croat. Chem. Acta, 77:331–344, 2004; and references cited therein.

[2] S. L. Taraviras, O. Ivanciuc, and D. Cabrol-Bass. *Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach.* J. Chem. Inf. Comput. Sci., 40:1128–1146, 2000; and references cited therein.

[3] S. Nikolić, G. Kovačević, A. Miličević, and N. Trinajstić. *The Zagreb Indices 30 Years After.* Croat. Chem. Acta, 76:113–127, 2003.

[4] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN–QSPR, a Software Package for the Search of Quantitative Structure Property Relationships.* MATCH Commun. Math. Comput. Chem., 51:187–204, 2004.

[5] J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN–CID, A Canonizer for Molecules and Graphs Accessible through the Internet.* J. Chem. Inf. Comput. Sci., 44:542–548, 2004.

[6] M. Meringer. *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation.* PhD thesis, Universität Bayreuth, 2004 (in German).

[7] N. L. Allinger. *MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms.* J. Am. Chem. Soc., 99:8127–8134, 1977.

[8] L. B. Kier and L. H. Hall. *Molecular Structure Description. The Electrotopological State.* Academic Press, San Diego, London, 1999.

[9] B. Ren. *Atomic–Level–Based AI Topological Descriptors for Structure–Property Correlations.* J. Chem. Inf. Comput. Sci., 43:161–169, 2003.

[10] D. Bonchev and N. Trinajstić. *Overall Molecular Descriptors. 3. Overall Zagreb Indices.* SAR QSAR Environ. Res., 12:213–236, 2001.

[11] D. Bonchev. *The Overall Wiener Index — A New Tool for Characterization of Molecular Topology.* J. Chem. Inf. Comput. Sci., 41:582–592, 2001.

[12] S. A. Wildman and G. M. Crippen. *Prediction of Physicochemical Parameters by Atomic Contributions.* J. Chem. Inf. Comput. Sci., 39:868–873, 1999.

[13] R. Todeschini and V. Consonni. *Handbook of Molecular Descriptors.* Wiley–VCH, Weinheim, 2000.

[14] C. Rücker, J. Braun, A. Kerber, and R. Laue. *The Molecular Descriptors Computed with MOLGEN.* http://www.mathe2.uni-bayreuth.de/molgenqspr, 2003.

[15] I. Gutman and K. C. Das. *The First Zagreb Index 30 Years After.* MATCH Commun. Math. Comput. Chem., 50:83–92, 2004.

[16] I. Gutman, B. Ruščić, N. Trinajstić, and C. F. Wilcox Jr. *Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes.* J. Chem. Phys., 62:3399–3405, 1975.

[17] C. Rücker and G. Rücker. *Counts of All Walks as Atomic and Molecular Descriptors.* J. Chem. Inf. Comput. Sci., 33:683–695, 1993.

[18] C. Rücker and G. Rücker. *Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules.* J. Chem. Inf. Comput. Sci., 40:99–106, 2000.

[19] I. Gutman, C. Rücker, and G. Rücker. *On Walks in Molecular Graphs.* J. Chem. Inf. Comput. Sci., 41:739–745, 2001.