MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **79** (2018) 5-54

ISSN 0340 - 6253

New Geometric-Arithmetic Indices

Piotr Wilczek

Computer Laboratory, Na Skarpie 99/24, 61-163 Poznań, POLAND
piotr.wilczek.net@onet.pl

(Received May 22, 2017)

Abstract

Based on the definition of the general geometric-arithmetic index, this article introduces several new geometric-arithmetic indices (Part One). In Part Three, we have determined the degree of degeneracy of these new invariants and we have designated the first pairs (or subsets) of molecular graphs having the same values of a given geometric-arithmetic index. It appears that in many cases the newly proposed descriptors have a much greater level of uniqueness that the strongly discriminating Balaban J index. In Part Four, we have demonstrated the applicability of these newly defined molecular descriptors for QSPR studies. Namely, we have used them to model certain physicochemical properties of several classes of organic compounds. The results of internal and external validations of the obtained models have indicated that the models based on the new geometric-arithmetic indices have high descriptive and predictive capabilities and are externally stable. Also, these results have testified that the QSPR models based on these new topological indices in many cases outperform models known in the literature. Therefore, it can be speculated that these new geometric-arithmetic indices will be used in future QSPR/QSAR studies.

1. Introduction

One of the fundamental topics in all quantitative structure property/activity relationship (QSPR/QSAR) studies is the transformation of chemical structures into molecular invariants which, in turn, should be correlated with certain specific physicochemical properties or biological (or toxicological/pharmacological) activities. Consequently, it is of primary importance for any future QSPR/QSAR investigations to search for novel highly correlating and highly discriminating molecular descriptors.

Let $G = \big(V(G), E(G)\big)$ denote a molecular graph where $V(G) = \{v_1, v_2, \dots, v_n\}$ is the vertex set and $E(G)$ is the edge set. The topological distance between two vertices $v_i, v_j \in V(G)$, denoted by $d_G\big(v_i, v_j\big)$, is identified with the number of edges in any shortest path connecting them. The eccentricity $\varepsilon_G(v_i)$ of a vertex $v_i \in V(G)$ is the greatest topological distance

between v_i and any other vertex in G. The diameter of a molecular graph G, denoted by $diam(G)$, is defined as $diam(G) = \max_{v_i \in V(G)} \varepsilon_G(v_i)$. The symbol $deg_G(v_i)$ denotes the degree (i.e., the number of first neighbors) of the vertex $v_i \in V(G)$. For two vertices $v_i, v_j \in V(G)$, $v_i v_j$ means that v_i and v_j are adjacent, i.e., $v_i v_j \in E(G)$.

In recent years, a whole novel family of topological invariants has been introduced [11]. These new descriptors are termed as the "*geometric-arithmetic indices*" and their formal definition can be expressed as follows:

$$GA_{general} = GA_{general}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{f(v_i) f(v_j)}}{\frac{1}{2}\left(f(v_i) + f(v_j)\right)}$$

where $v_i, v_j \in V(G)$ and $f(v_i)$ is some quantity that can be uniquely connected with the vertex v_i of a molecular graph G. The *first geometric-arithmetic index* (GA_1) was suggested by D. Vukičević and B. Furtula by postulating $f(v_i)$ to be the degree ($deg_G(v_i)$) of the vertex $v_i \in V(G)$ [49]. Hence, this descriptor has the form:

$$GA_1(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{deg_G(v_i) deg_G(v_j)}}{\frac{1}{2}\left(deg_G(v_i) + deg_G(v_j)\right)}.$$

To present further geometric-arithmetic descriptors, let us recall the subsequent terminology: for any edge $v_i v_j \in E(G)$, let us define the following two quantities:

$$n_{v_i} = \left|\{x \in V(G) | d_G(x, v_i) < d_G(x, v_j)\}\right|$$

and

$$n_{v_j} = \left|\{x \in V(G) | d_G(x, v_i) > d_G(x, v_j)\}\right|.$$

Thus, n_{v_i} is equal to the number of vertices of the molecular graph G which are located closer to $v_i \in V(G)$ than to $v_j \in V(G)$. On the other hand, the quantity n_{v_j} is equal to the number of vertices of the molecular graph G which are located closer to $v_j \in V(G)$ than to $v_i \in V(G)$ [13]. Then, the *second geometric-arithmetic index*, introduced by G. Fath-Tabar et al. [19], can be expressed as follows:

$$GA_2(G) = \sum_{v_iv_j \in E(G)} \frac{\sqrt{n_{v_i} n_{v_j}}}{\frac{1}{2}\left(n_{v_i} + n_{v_j}\right)}.$$

Suppose that $h = st$ is an edge linking two vertices $s, t \in V(G)$. The distance between any vertex $v_i \in V(G)$ and the edge h in the molecular graph G is defined as: $d_G(v_i, h) = min\{d_G(v_i, s), d_G(v_i, t)\}$. Then, the subsequent two quantities:

$$m_{v_i} = \left|\{h \in E(G) | d_G(h, v_i) < d_G(h, v_j)\}\right|$$

and

$$m_{v_j} = \left|\{h \in E(G) | d_G(h, v_i) > d_G(h, v_j)\}\right|$$

correspond to the number of edges of the molecular graph G which are located closer to $v_i \in V(G)$ than to $v_j \in V(G)$ and to the number of edges of the molecular graph G which are situated closer to $v_j \in V(G)$ than to $v_i \in V(G)$, respectively [13]. Now, the *third geometric-arithmetic index*, defined by B. Zhou et. al. [50], has the form:

$$GA_3(G) = \sum_{v_iv_j \in E(G)} \frac{\sqrt{m_{v_i} m_{v_j}}}{\frac{1}{2}\left(m_{v_i} + m_{v_j}\right)}.$$

Later on, M. Ghorbani et al. [21] suggested the *fourth geometric-arithmetic* index whose formula is as follows:

$$GA_4(G) = \sum_{v_iv_j \in E(G)} \frac{\sqrt{\varepsilon_G(v_i)\varepsilon_G(v_j)}}{\frac{1}{2}\left(\varepsilon_G(v_i) + \varepsilon_G(v_j)\right)}$$

and A. Graovac et al. [22] considered the *fifth geometric-arithmetic* descriptor of the form:

$$GA_4(G) = \sum_{v_iv_j \in E(G)} \frac{\sqrt{\delta_G(v_i)\delta_G(v_j)}}{\frac{1}{2}\left(\delta_G(v_i) + \delta_G(v_j)\right)}$$

where $\delta_G(v_i) = \sum_{v_iu \in E(G)} deg_G(u)$.

Also, the edge and total versions of geometric-arithmetic index were introduced [33, 34].

Some mathematical properties (e.g., lower and upper bounds, extremal graphs, inequalities, Nordhaus-Gaddum-type results, spectral characteristic) of these indices are treated in [8, 9, 10, 11, 12, 13, 44, 45].

Also, it was demonstrated that GA_1, GA_2 and GA_3 descriptors possess relatively good descriptive as well as predictive capabilities with respect to some selected properties of octanes and benzenoid hydrocarbons [11, 49]. The degeneracy of GA_1, GA_2 and GA_3 indices was studied in [14].

The contribution of the present report is fourfold. Firstly, we will define 9 new geometric-arithmetic indices. Secondly, we will determine the degree of degeneracy of these newly proposed invariants. Thirdly, we will designate minimal pairs (or subsets) of molecular graphs having the same values of a given geometric-arithmetic index. Fourthly, we will build several representative QSPR models to demonstrate the usefulness for chemical research of these newly defined molecular descriptors.

To formally introduce these novel descriptors, let us recall the following terminology [24, 26, 27, 48]: the *distance matrix* of any molecular graph $G = \big(V(G), E(G)\big)$ where $|V(G)| = n$, denoted by $D(G)$, is a real symmetric $n \times n$ matrix whose entries $[D]_{ij}$ correspond to the topological distance between the vertices $v_i, v_j \in V(G)$, the *Harary matrix* (also known as the *reciprocal distance matrix*) of any molecular graph G with n vertices, denoted by $RD(G)$, is a real symmetric $n \times n$ matrix whose elements $[RD]_{ij}$ are given by $[RD]_{ij} = \frac{1}{d_G(v_i, v_j)}$ if $v_i \neq v_j$ and $[RD]_{ij} = 0$ otherwise. On the other hand, the *reverse Wiener matrix* (also known as the *reverse distance matrix*) of any molecular graph G where $|V(G)| = n$, denoted by $RW(G)$, is a real symmetric $n \times n$ matrix whose entries $[RW]_{ij}$ are given by $[RW]_{ij} = diam(G) - d_G\big(v_i, v_j\big)$ if $v_i \neq v_j$ and $[RW]_{ij} = 0$ otherwise. The *Randić matrix* (also known as the *product connectivity matrix*) of a molecular graph G with n vertices, denoted by $\chi(G)$, is identified with a real symmetric $n \times n$ matrix whose elements $[\chi]_{ij}$ are given by $[\chi]_{ij} = \Big(deg_G(v_i) deg_G\big(v_j\big)\Big)^{-\frac{1}{2}}$ if $v_i \neq v_j$ and $[\chi]_{ij} = 0$ otherwise.

For any molecular matrix $M(G)$ associated with a molecular graph $G = \big(V(G), E(G)\big)$, the *Vertex Sum operator* (also known as the *Row Sum operator*) for the vertex $v_i \in V(G)$,

denoted by $VS\big(M(G)\big)_i$, is defined as the sum of the entries in the row i of the graph-theoretical matrix $M(G)$, i.e.,

$$VS\big(M(G)\big)_i = \sum_{j=1}^{n} [M(G)]_{ij}.$$

If $M(G)$ is the distance matrix $D(G)$, then the operator $VS\big(M(G)\big)_i$ gives the *distance sum* of the vertex $v_i \in V(G)$. If $M(G)$ is the Harary matrix $RD(G)$, then the operator $VS\big(M(G)\big)_i$ gives the *reciprocal distance sum* corresponding to the vertex $v_i \in V(G)$ and if $M(G)$ is the reverse Wiener matrix $RW(G)$, then the operator $VS\big(M(G)\big)_i$ produces the *reverse distance sum* of the vertex $v_i \in V(G)$ [25, 48].

Consequently, it can be easily observed that for any molecular graph G and any vertex $v_i \in V(G)$ it is possible to define the following vertex invariants: $VS\big(D(G)\big)_i$, $VS\big(RD(G)\big)_i$, $VS\big(RW(G)\big)_i$ and $VS\big(\chi(G)\big)_i$. These quantities correspond to the row sums of the distance matrix, the reciprocal distance matrix, the reverse Wiener matrix and the product connectivity matrix associated with the molecular graph G.

Based on the notion of the distance sum of a vertex $v_i \in V(G)$, A. A. Dobrynin and A. A. Kochetova introduced the so-called *degree distance* of $v_i \in V(G)$ [15]. For any molecular graph $G = \big(V(G), E(G)\big)$ and any vertex $v_i \in V(G)$, this novel vertex invariant, denoted by $D'(v_i)$, is defined as follows:

$$D'(v_i) = deg_G(v_i) VS\big(D(G)\big)_i.$$

Other quantities which can be uniquely connected with a vertex $v_i \in V(G)$ include the so-called *centrality measures*. Such measures determine the most important elements in a given graph G. They are mainly studied in the field of *Social Network Analysis*. In the present article, we will be concerned with such centrality metrics as the eigenvector centrality, the parameterized exponential subgraph centrality, the parameterized total subgraph communicability, the resolvent subgraph centrality as well as with the Katz centrality.

The *eigenvector centrality* EC_i of any vertex $v_i \in V(G)$ is identified with the i-th component of the eigenvector associated with the largest eigenvalue of the adjacency matrix $A(G)$ of G, i.e.,

$$EC_i = \boldsymbol{q}_1(i)$$

where $\boldsymbol{q}_1$ is the dominant eigenvector of $A(G)$ [48]. A vertex $v_i \in V(G)$ possesses the high value of the eigenvector centrality if it is adjacent to many other vertices or if it is linked to other nodes that themselves have high value of this centrality measure.

Such centralities as the *parameterized exponential subgraph centrality* and the *parameterized total subgraph communicability* are based on the notion of the *parameterized matrix exponential* which is defined for any molecular graph G by the condition

$$e^{\beta A(G)}$$

where $A(G)$ is the adjacency matrix connected with G and $\beta > 0$ [3, 16, 30]. The eigenvalues of $e^{\beta A(G)}$ are given by $e^{\beta\lambda_1}$, $e^{\beta\lambda_2}$, …, $e^{\beta\lambda_n}$ where λ_1, λ_2, …, λ_n are the eigenvalues of the adjacency matrix $A(G)$. The power series expansion of $e^{\beta A(G)}$ is given by

$$e^{\beta A(G)} = I + \beta A(G) + \frac{\beta^2\left(A(G)\right)^2}{2!} + \ldots + \frac{\beta^k\left(A(G)\right)^k}{k!} + \ldots = \sum_{k=0}^{\infty}\frac{\beta^k\left(A(G)\right)^k}{k!}.$$

The parameterized exponential subgraph centrality of a vertex $v_i \in V(G)$ is given by

$$SC_i(\beta) = \left[e^{\beta A(G)}\right]_{ii}.$$

It is well known in the field of *Graph Theory* that if $A(G)$ is the adjacency matrix of a graph G, then the entry $\left(A(G)\right)_{ij}^k$ is equal to the number of walks of length k between two vertices $v_i, v_j \in V(G)$. Recall that a *walk* of length k defined on a molecular graph $G = \left(V(G), E(G)\right)$ is identified with a sequence of vertices $v_i, v_2, \ldots, v_{k+1}$ such that $v_i v_{i+1} \in E(G)$ for all $1 \leq i \leq k$. A *closed walk* is identified with a walk that begins and ends at the same node. Consequently, it follows that the exponential subgraph centrality of a vertex $v_i \in V(G)$ (which is equal to $\left[e^{\beta A(G)}\right]_{ii}$) identifies the number of closed walks centered at v_i. This centrality metrics weights a walk of length equal to k by a factor $\frac{\beta^k}{k!}$. Roughly speaking, the exponential subgraph centrality estimates the number of subgraphs a node $v_i \in V(G)$ participates in, weighting them with respect to their size.

On the other hand, the quantity $\left[e^{\beta A(G)}\right]_{ij}$ characterizes the communicability between the vertices v_i and v_j in any molecular graph G. Therefore, the row sum of the matrix $e^{\beta A(G)}$ for a vertex $v_i \in V(G)$ given by

$$TC_i(\beta) = VS\left(e^{\beta A(G)}\right)_i = \sum_{j=1}^{n}\left[e^{\beta A(G)}\right]_{ij}$$

identifies all walks between the vertex v_i and all other vertices in the graph G (including the vertex v_i). In this context, the quantity $TC_i(\beta)$ is referred to as the *parameterized total subgraph communicability* of the node $v_i \in V(G)$. This centrality weights walks of length equal to k by a factor $\frac{\beta^k}{k!}$ [3, 30].

The above-listed two centrality measures which are defined in terms of the diagonal entries or the row sums of the parameterized exponential of the adjacency matrix of any graph G (i.e., $e^{\beta A(G)}$) were successfully used, for instance, in protein biochemistry (e.g., the identification of crucial proteins in proteomic maps) [17, 18], pathophysiology (e.g., the description of malignant tissues)[39] and neurophysiology (e.g., the characterization of healthy and stroke-damaged brain networks) [6].

The second class of centrality measures whose formal definitions are also expressed in terms of the matrix function $f\left(A(G)\right)$ where $A(G)$ is the adjacency matrix linked with a molecular graph G are the so-called matrix resolvent-based centrality metrics. Recall that a *matrix resolvent* of $A(G)$ for a molecular graph $G = \left(V(G), E(G)\right)$ is given by $\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}$ where $\boldsymbol{I}$ is the $n\times n$ identity matrix and $0 < \alpha < \frac{1}{\lambda_1}$. Here, λ_1 denotes the spectral radius of the adjacency matrix $A(G)$ [3, 30]. This matrix function possesses eigenvalues of the form $\frac{1}{1-\alpha\lambda_i}$ where λ_i are the eigenvalues of the adjacency matrix $A(G)$. The power series expansion of $\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}$ is given by:

$$\left(\boldsymbol{I} - \alpha A(G)\right)^{-1} = \boldsymbol{I} + \alpha A(G) + \alpha^2\left(A(G)\right)^2 + \ ... + \alpha^k\left(A(G)\right)^k + \ ... = \sum_{k=0}^{\infty}\alpha^k\left(A(G)\right)^k.$$

The constraints imposed on the value of α (i.e., $0 < \alpha < \frac{1}{\lambda_1}$) imply that the matrix $\boldsymbol{I} - \alpha A(G)$ is invertible and that the above geometric series is convergent to its inverse. Such selection of α also implies that the matrix $\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}$ is non-negative. Based on the notion of the

matrix resolvent $\left(\boldsymbol{I}-\alpha A(G)\right)^{-1}$ it is possible to formulate the following two definitions [3, 30]: the *resolvent subgraph centrality* for a vertex $v_i \in V(G)$ of a molecular graph $G = \left(V(G), E(G)\right)$, denoted by $RC_i(\alpha)$, is equal to the diagonal entries of the matrix resolvent of the adjacency matrix $A(G)$, i.e.,

$$RC_i(\alpha) = \left[\left(\boldsymbol{I}-\alpha A(G)\right)^{-1}\right]_{ii};$$

the *Katz centrality* for a vertex $v_i \in V(G)$ of a molecular graph $G = \left(V(G), E(G)\right)$, denoted by $K_i(\alpha)$, is equal to the row sums of the matrix resolvent of the adjacency matrix $A(G)$, i.e.,

$$K_i(\alpha) = VS\left(\left(\boldsymbol{I}-\alpha A(G)\right)^{-1}\right)_i = \sum_{j=1}^{n}\left[\left(\boldsymbol{I}-\alpha A(G)\right)^{-1}\right]_i.$$

The first centrality metrics $RC_i(\alpha)$ identifies the number of closed walks centered at the vertex $v_i \in V(G)$ whereas the second centrality metrics $K_i(\alpha)$ identifies the total number of walks between the vertex $v_i \in V(G)$ and all other vertices in the molecular graph G. Both measures weights walks of length equal to k by α^k.

Thus, we have obtained for any molecular graph $G = \left(V(G), E(G)\right)$ and any node $v_i \in V(G)$ the following vertex invariants: $VS\left(D(G)\right)_i$, $VS\left(RD(G)\right)_i$, $VS\left(RW(G)\right)_i$, $VS\left(\chi(G)\right)_i$, $D'(v_i)$, EC_i, $SC_i(\beta)$, $TC_i(\beta)$, $RC_i(\alpha)$ and $K_i(\alpha)$.

Based on the above considerations and on the definition of the general geometric-arithmetic index, the following definition seems to be justified:

$$GA_6(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS\left(M(G)\right)_i VS\left(M(G)\right)_j}}{\frac{1}{2}\left(VS\left(M(G)\right)_i + VS\left(M(G)\right)_j\right)}$$

where $M(G)$ is any molecular matrix associated with G. Thus, in the case of the sixth geometric-arithmetic index ($GA_6(G)$), the quantity $f(v_i)$ uniquely connected with $v_i \in V(G)$ is identified with the row sum of $M(G)$ corresponding to the vertex v_i[1]. Consequently, in this paper, we will single out the following subtypes of the sixth geometric-arithmetic index:

[1] Undoubtedly, if $M(G)$ is an adjacency matrix then we obtain the first geometric-arithmetic index. But when $M(G)$ is not an adjacency matrix then we will get new interesting topological indices.

$$GA_{6a}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS(D(G))_i VS(D(G))_j}}{\frac{1}{2}\left(VS(D(G))_i + VS(D(G))_j\right)},$$

$$GA_{6b}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS(RD(G))_i VS(RD(G))_j}}{\frac{1}{2}\left(VS(RD(G))_i + VS(RD(G))_j\right)},$$

$$GA_{6c}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS(RW(G))_i VS(RW(G))_j}}{\frac{1}{2}\left(VS(RW(G))_i + VS(RW(G))_j\right)},$$

$$GA_{6d}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS(\chi(G))_i VS(\chi(G))_j}}{\frac{1}{2}\left(VS(\chi(G))_i + VS(\chi(G))_j\right)},$$

$$GA_{6e}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{TC_i(\beta) TC_j(\beta)}}{\frac{1}{2}\left(TC_i(\beta) + TC_j(\beta)\right)} = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS(e^{\beta A(G)})_i VS(e^{\beta A(G)})_j}}{\frac{1}{2}\left(VS(e^{\beta A(G)})_i + VS(e^{\beta A(G)})_j\right)}$$

for $\beta > 0$,

$$GA_{6f}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{K_i(\alpha) K_j(\alpha)}}{\frac{1}{2}\left(K_i(\alpha) + K_j(\alpha)\right)} = \sum_{v_i v_j \in E(G)} \frac{\sqrt{VS\left(\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}\right)_i VS\left(\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}\right)_j}}{\frac{1}{2}\left(VS\left(\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}\right)_i + VS\left(\left(\boldsymbol{I} - \alpha A(G)\right)^{-1}\right)_j\right)}$$

for $0 < \alpha < \frac{1}{\lambda_1}$ where λ_1 is the spectral radius of $A(G)$.

In the above cases, the quantity $f(v_i)$ uniquely associated with $v_i \in V(G)$ is identified with the row sums (corresponding to v_i) of the following molecular matrices: the distance matrix ($GA_{6a}(G)$), the reciprocal distance matrix ($GA_{6b}(G)$), the reverse Wiener matrix ($GA_{6c}(G)$), the Randić matrix ($GA_{6d}(G)$), the parameterized matrix exponential of $A(G)$ (i.e., the parameterized total subgraph communicability of the node $v_i \in V(G)$) ($GA_{6e}(G)$), the matrix resolvent of $A(G)$ (i.e., the Katz centrality of the node $v_i \in V(G)$) ($GA_{6f}(G)$).

In the case of molecular matrices without zeros on the main diagonal, the following general geometric-arithmetic index is proposed:

$$GA_7(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{[M(G)]_{ii}[M(G)]_{jj}}}{\frac{1}{2}\left([M(G)]_{ii} + [M(G)]_{jj}\right)}.$$

.

In the above expression $[M(G)]_{ii}$ denotes the diagonal element corresponding to the vertex $v_i \in V(G)$. Therefore, in this work, we will single out the following subtypes of the seventh geometric-arithmetic index:

$$GA_{7a}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{SC_i(\beta) SC_j(\beta)}}{\frac{1}{2}\left(SC_i(\beta) + SC_j(\beta)\right)} = \sum_{v_i v_j \in E(G)} \frac{\sqrt{[e^{\beta A(G)}]_{ii}[e^{\beta A(G)}]_{jj}}}{\frac{1}{2}\left([e^{\beta A(G)}]_{ii} + [e^{\beta A(G)}]_{jj}\right)}$$

for $\beta > 0$,

$$GA_{7b}(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{RC_i(\alpha) RC_j(\alpha)}}{\frac{1}{2}\left(RC_i(\alpha) + RC_j(\alpha)\right)} = \sum_{v_i v_j \in E(G)} \frac{\sqrt{\left[\left(I - \alpha A(G)\right)^{-1}\right]_{ii}\left[\left(I - \alpha A(G)\right)^{-1}\right]_{jj}}}{\frac{1}{2}\left(\left[\left(I - \alpha A(G)\right)^{-1}\right]_{ii} + \left[\left(I - \alpha A(G)\right)^{-1}\right]_{jj}\right)}$$

for $0 < \alpha < \frac{1}{\lambda_1}$ where λ_1 is the spectral radius of $A(G)$.

Thus, in the cases of these subtypes, the quantity $f(v_i)$ uniquely associated with $v_i \in V(G)$ is identified with the parameterized exponential subgraph centrality of the node v_i ($GA_{7a}(G)$) or with the resolvent subgraph centrality of the node v_i ($GA_{7b}(G)$).

Also, it seems possible to introduce the eighth geometric-arithmetic index as well as the ninth geometric-arithmetic index. Their formal definitions are listed below:

$$GA_8(G) = \sum_{v_i v_j \in E(G)} \frac{\sqrt{D'(v_i) D'(v_j)}}{\frac{1}{2}\left(D'(v_i) + D'(v_j)\right)},$$

$$GA_9(G) = \sum_{v_i v_j \in E(G)} \sum_{v_i v_j \in E(G)} \frac{\sqrt{EC_i EC_j}}{\frac{1}{2}\left(EC_i + EC_j\right)}.$$

In these cases, the quantity $f(v_i)$ uniquely connected with $v_i \in V(G)$ is given by the degree distance of the vertex v_i ($GA_8(G)$) or by the eigenvector centrality of the vertex v_i ($GA_9(G)$).

2. Datasets and computational methods

All numerical experiments were performed on a synthetic dataset of all exhaustively generated non-isomorphic, undirected and connected graphs having up to 7 vertices with the exception of the unique graph with $|V(G)| = 1$ and $|E(G)| = 0$. This dataset, denoted by $\mathcal{G}$,

contains 995 graphs (1 graph with $|V(G)| = 2$, 2 graphs with $|V(G)| = 3$, 6 graphs with $|V(G)| = 4$, 21 graphs with $|V(G)| = 5$, 112 graphs with $|V(G)| = 6$ and 853 graphs with $|V(G)| = 7$). These quantities are in agreement with the Pólya enumeration theory [40]. Graphs from the dataset $\mathcal{G}$ are numbered from 1 (the graph K_2) to 995 (graph K_7).

In order to quantitatively assess the *uniqueness* (i.e., the *degree of degeneracy*) of a particular molecular descriptor TI, the *sensitivity index* $S(TI)$ introduced by E. V. Konstantinova was used [31]. This index is defined as

$$S(TI) = \frac{|\mathcal{G}| - |degen(\mathcal{G})|}{|\mathcal{G}|}$$

where $|\mathcal{G}|$ denotes the cardinality of any graph dataset $\mathcal{G}$ on which TI was tested (in our case $|\mathcal{G}| = 995$) and $|degen(\mathcal{G})|$ is equal to the number of degeneracies of TI within $\mathcal{G}$. It is immediately apparent that when $S(TI) = 1$, then the analyzed graph dataset $\mathcal{G}$ does not contain any pair of non-isomorphic graphs with the same values of TI. Also, it can be easily demonstrated that the sensitivity index $S(TI)$ of a given topological descriptor TI is dependent on the selected decimal places. Consequently, in this work all molecular invariants were calculated with an accuracy of 9 decimal places.

When calculating GA_9 index, the eigenvector centrality is scaled so that the maximum score is equal to 1.

The publicly available dataset of octane isomers was downloaded from the webpage www.moleculardescriptors.eu. The dataset of 39 saturated alkanes with their experimental boiling points (Table 12) was taken from [42]. The dataset of 29 aliphatic alcohols with their experimental enthalpies of combustion (Table 15) was borrowed from [4, 20, 36, 38, 47]. The dataset of 42 aliphatic alcohols with their experimental molar volumes (Table 20) was taken from [37]. The dataset of 41 aliphatic alcohols with their experimental molar refractions (Table 20) was also borrowed from [37]. The dataset consisting of 22 aldehydes and 24 ketones with their experimental molar refractions (Table 23) was taken from [43]. The dataset composed of 3 aldehydes and 15 ketones with their experimental gas heat capacities (Table 26) was also borrowed from [43]. The dataset of 20 monocarboxylic acids with their experimental enthalpies of formation and combustion (Table 31) was taken from [1, 32, 38, 46].

In order to establish QSPR models, Linear Regression (simple and multiple) as well as Power Regression have been used [5]. To monitor the *descriptive* capabilities (i.e., the goodness of fit) of the obtained regression equations, the *correlation coefficient* (r), the *coefficient of determination* (R^2), the *standard deviation* (s) and the *Fisher ratio* (F) were utilized as statistical parameters [5, 29, 48]. As postulated by Z. Mihalić and N. Trinajstić [35], a good QSPR model must have a value of $r > 0.99$ while the values of s depend on the property under study. To test the *predictive* capabilities (i.e., the goodness of prediction) of the obtained regression models, the *leave-one-out procedure of cross-validation* (Q^2) was employed. Also, the *standard deviation error in prediction* (SDEP) was calculated [29, 48]. To exclude the possibility of chance correlation, the *y-randomization* (*y-scrambling*) test was used. As suggested in [29], if $R^2_{yrand} < 0.2$ and $Q^2_{yrand} < 0.2$, then there is no chance correlation. Here, R^2_{yrand} and Q^2_{yrand} stand for the basic statistics of the randomized models. For each final regression equation, the y-randomization test was repeated 1000 times. To verify if the obtained model has satisfactory predictive abilities with respect to external data, we used a procedure in which the entire dataset is randomly divided into three subsets (A, B or C) and each subset (A or B or C) is predicted by using the other two subsets (BC or AC or AB) as the training set [28]. The quality of fit between the predicted values and the experimental data was monitored by the values of R^2 and s.

All simulations and computations contained in the following paper were conducted in the R programming language [7, 23, 41].

3. Degeneracy of the geometric-arithmetic indices

It is well known that when two molecular graphs are *topologically* identical, i.e., isomorphic, then they also possess identical values of all graph invariants. Although, the reverse correspondence is not universally true. This means that the identical values of any given graph descriptors do not imply the isomorphism of the molecular graphs. Generally speaking, any topological descriptor is said to be *degenerate* when there exist at least two non-isomorphic graphs having identical values of that invariant. The uniqueness of the graph-theoretical descriptors has been studied many times in the field of computational chemistry. For instance, it has been observed that the level of degeneracy is high for the Wiener index (W), the Harary index (H), the Hosoya index (Z) and the Zagreb indices ($Z_1$and Z_2), lower for

the Randić connectivity index (χ) and very low for the Balaban J index [2, 14]. In order to diminish the degree of degeneracy of first- and second-generation molecular descriptors, D. Bonchev and N. Trinajstić developed the so-called information-theoretical indices [2].

To sum up, it can be uttered that the discriminative power is one of the fundamental properties of each topological index. This characteristic quantitatively evaluates the capability of molecular descriptors to distinguish non-isomorphic chemical graphs.

In the present section, our main aim is to scrutinize the extent to which the newly introduced topological indices are degenerate as well as find the smallest pairs (or subsets) of graphs for which the given geometric-arithmetic descriptor has the same value. The following studies are carried out on the dataset $\mathcal{G}$ of all exhaustively generated non-isomorphic, undirected and connected graphs having from 2 to 7 vertices. Table 1 presents the values of the sensitivity index for GA_1, GA_4, GA_{6a}, GA_{6b}, GA_{6c}, GA_{6d}, GA_{6e}, GA_{6f}, GA_{7a}, GA_{7b}, GA_8 and GA_9 indices, the first pairs (or subsets) of the graphs from $\mathcal{G}$ having the same value of the given geometric-arithmetic index as well as the values of that index for those minimal indistinguishable graphs. The graphs from Table 1 are shown in Figure 1. To make our results more illustrative, let us recall that the values of the sensitivity index $S(TI)$ evaluated on the same dataset $\mathcal{G}$ for the aforementioned topological descriptors are equal to $S(W) = 0.011$, $S(H) = 0.043$, $S(Z) = 0.017$, $S(Z_1) = 0.015$, $S(Z_2) = 0.103$, $S(\chi) = 0.472$ and $S(J) = 0.83$, respectively.

Table 1. Degeneracy of twelve geometric-arithmetic indices.

Index	$\lvert degen(\mathcal{G})\rvert$	$S(GA)$	The first pair (or subset) of the graphs from $\mathcal{G}$ having the same value of GA	The value of GA for the graphs from the fourth column
GA_1	448	0.550	43, 45	5.691642602
GA_4	926	0.069	8, 13	4.771236166
GA_{6a}	173	0.826	9, 49	6
GA_{6b}	171	0.828	9, 49	6
GA_{6c}	179	0.820	28, 88, 90	7.958973274
GA_{6d}	70	0.930	9, 49	6
GA_{6e} (β=0.005)	177	0.822	21, 46	5.999987716
GA_{6e} (β=0.03)	50	0.95	9, 49	6

GA_{6e} (0.055≤β≤0.105)	46	0.954	9, 49	6
GA_{6e} (0.13≤β≤9.98)	42	0.958	9, 49	6
GA_{6f} (α=0.0025)	98	0.902	9, 49	6
GA_{6f} (α=0.005)	59	0.941	9, 49	6
GA_{6f} (0.0075≤α≤0.0175)	46-48	0.954-0.952	9, 49	6
GA_{6f} (0.02≤α≤0.0325 and α=0.7725)	44	0.956	9, 49	6
GA_{6f} (0.035≤α≤0.77 and 0.775≤α≤0.9975)	42	0.958	9, 49	6
GA_{7a} (β=0.005)	983	0.012	3, 4, 5	3
GA_{7a} (β=0.03)	219	0.78	43, 45, 151	5.999999747
GA_{7a} (β=0.055)	49	0.951	22, 48	5.999998292
GA_{7a} (β=0.08)	12	0.988	22, 48	5.999992393
GA_{7a} (0.105≤β≤3.78)	4	0.996	9, 49	6
GA_{7a} (3.805≤β≤7.58)	6	0.994	9, 49	6
GA_{7a} (7.605≤β≤9.98)	8-14	0.992-0.986	9, 49	6
GA_{7b} (0.0025≤α≤0.3275)	4-991	0.004-0.996	Many different pairs/subsets of graphs	-
GA_{7b} (0.33≤α≤0.9975)	4	0.996	9, 49	6
GA_8	169	0.830	9, 49	6
GA_9	42	0.958	9, 49	6

From Table 1, it can be seen that the degree of degeneracy of GA_4 index is very high, lower for GA_1 and very low for GA_{6a}, GA_{6b}, GA_{6c}, GA_{6d}, GA_8 and GA_9 indices. It was hypothesized

that the degree of uniqueness of the geometric-arithmetic indices whose formulae include the adjustable parameters β or α strongly depends on these parameters. Therefore, in our computational studies, in the case of GA_{6e} and GA_{7a} indices, the adjustable parameter β had the form of a sequence of real numbers from 0.005 to 9.98 with an increment equal to 0.025. On the other hand, in the case of GA_{6f} and GA_{7b} indices, the adjustable parameter α had the form of a sequence of real numbers from 0.0025 to 0.9975 with an increment equal to 0.0025. The relationship between the adjustable parameters (β and α) and the values of $S(TI)$ where $TI \in \{GA_{6e}, GA_{6f}, GA_{7a}, GA_{7b}\}$ is detailed in Figure 2. From this plot, it can be seen that the values of $S(GA_{6e})$ and $S(GA_{6f})$ range from 0.822 to 0.958 and from 0.902 to 0.958, respectively. On the other hand, GA_{7a} index has the very degenerate form (for β=0.005) as well as many forms with extremely low levels of degeneracy (for 0.08≤β≤9.98) whereas GA_{7b} index has several very degenerate forms as well as many forms with extremely low degrees of degeneracy.

Thus, in many cases we obtained the topological indices with very low level of degeneracy.

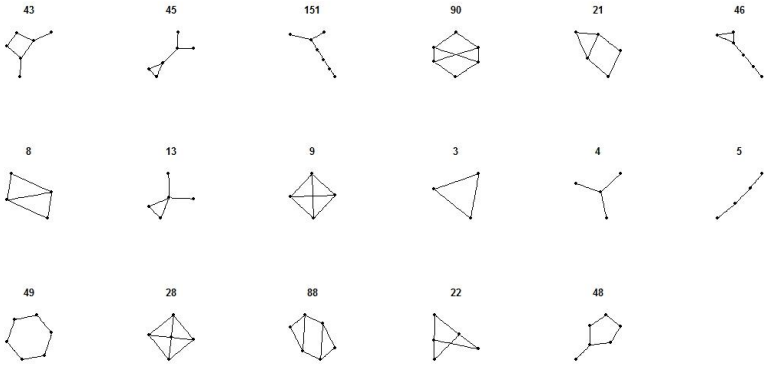

Figure 1. Graphs from Table 1.

Namely, note that the degree of degeneracy of GA_4 index is comparable to the degree of degeneracy of the Wiener index (or the Harary index or two Zagreb indices) whereas the level of degeneracy of GA_1 index is comparable to the level of degeneracy of the Randić index. On the other hand, the values of the sensitivity index for GA_{6a}, GA_{6b}, GA_{6c} and GA_8 indices are very close to $S(J)$ whereas GA_{6d}, GA_{6e}, GA_{6f}, GA_{7a}, GA_{7b} and GA_9 indices are significantly

less degenerate than the Balaban J index or have forms with a much higher level of uniqueness than J index.

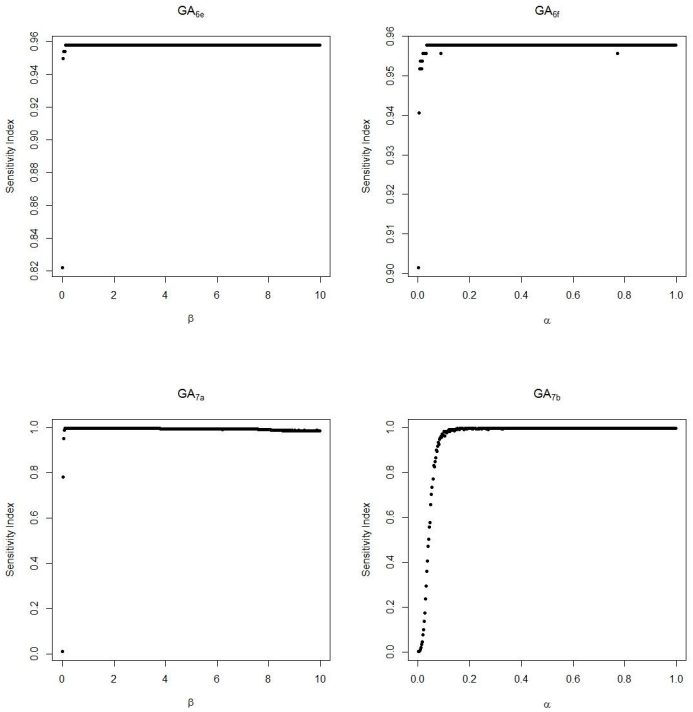

Figure 1. Correlations between adjustable parameters (β and α) and values of sensitivity index for four geometric-arithmetic descriptors.

4. Correlations to physical properties

It is widely recognized that topological descriptors based on molecular graphs can be easily computed using current computer techniques. Therefore, graph-theoretical approaches are often employed in QSAR/QSPR studies. In this section, we will demonstrate the applicability of the newly introduced topological descriptors in modelling certain physicochemical properties of several selected classes of organic compounds.

4.1 The dataset of octane isomers

Saturated alkanes constitute an especially attractive family of organic compounds which are often used as a starting point for any QSAR/QSPR investigations. One of the methodologies employed in such studies is to select a certain class of alkanes (for instance, C_8, C_9 or C_{10} isomers) in order to obtain comparable results. In our study, we have used the dataset of octane isomers. This reference dataset consists of 18 octane isomers and contains 16 physicochemical properties of these compounds. The dataset of octane isomers have been used repeatedly in QSAR/QSPR research and its use for any initial assessment of modelling properties of newly proposed topological descriptors is recommended by International Academy of Mathematical Chemistry. Note that using the dataset of octane isomers as some benchmark dataset it is possible to avoid the so-called size effect.

Table 2. Correlation coefficient (r) between twelve geometric-arithmetic indices and nine properties of octanes[1,2].

Index	Property								
	BP	S	DENS	HVAP	DHVAP	HFORM	ACENFAC	MON	MV
GA_1	**0.823**	**0.912**	-0.553	**0.941**	**0.966**	**0.858**	**0.912**	-0.777	0.538
GA_4	0.358	**0.804**	-0.601	0.616	0.674	0.314	**0.877**	-0.729	0.617
GA_{6a}	0.459	**0.923**	-0.739	0.691	0.784	0.483	**0.980**	**-0.905**	0.752
GA_{6b}	0.693	**0.954**	-0.640	**0.871**	**0.927**	0.721	**0.987**	**-0.922**	0.639
GA_{6c}	0.466	**0.886**	**-0.887**	0.621	0.716	0.498	**0.850**	**-0.813**	**0.880**
GA_{6d}	**0.905**	0.793	-0.423	**0.945**	**0.938**	**0.920**	0.771	-0.565	0.396
GA_{6e} (β)	0.783 (0.005)	**0.962** (0.555)	-0.671 (0.68)	**0.906** (0.005)	**0.939** (0.055)	**0.846** (0.005)	**0.996** (1.005)	**-0.943** (1.355)	0.673 (0.805)
GA_{6f} (α)	**0.835** (0.0025)	**0.959** (0.5875)	-0.660 (0.7775)	**0.937** (0.0025)	**0.955** (0.075)	**0.889** (0.0025)	**0.996** (0.805)	**-0.942** (0.865)	0.666 (0.82)
GA_{7a} (β)	0.784 (0.03)	**0.960** (1.93)	-0.647 (2.33)	**0.908** (0.555)	**0.944** (0.88)	**0.847** (0.005)	**0.995** (2.605)	**-0.928** (3.28)	0.652 (2.655)
GA_{7b} (α)	**0.884** (0.005)	**0.948** (0.93)	-0.645 (0.9725)	**0.957** (0.005)	**0.958** (0.51)	**0.930** (0.005)	**0.992** (0.97)	**-0.938** (0.9825)	0.653 (0.9775)
GA_8	**0.915**	0.773	-0.366	**0.947**	**0.931**	**0.946**	0.742	-0.526	0.338
GA_9	0.551	**0.906**	-0.624	0.757	**0.834**	0.563	**0.975**	**-0.922**	0.636

[1]The values of $|r|$ greater than 0.8 are in bold.

[2] In the case of GA_{6e}, GA_{6f}, GA_{7a}, GA_{7b} indices, the optimal values of the adjustable parameter (β or α) are in parentheses below the value of the correlation coefficient.

For the present study, we selected the following properties of octanes: the boiling point (BP), the entropy (S), the density (DENS), the enthalpy of vaporization (HVAP), the standard enthalpy of vaporization (DHVAP), the enthalpy of formation (HFORM), the acentric factor (ACENFAC), the motor octane number (MON) and the molar volume (MV). The reason for choosing these properties is that for this collection of physicochemical parameters at least one of the tested descriptors exhibits a relatively good linear correlation(i.e., $|r| > 0.8$). From Table 2, it can be seen that GA_{6f} and GA_{7b} indices exhibit relatively satisfactory correlations with seven properties of octanes, GA_1, GA_{6e} and GA_{7a} indices with six properties of octanes, GA_{6b} and GA_{6c} indices with five properties of octanes, GA_{6d}, GA_8 and GA_9 indices with four properties of octanes, GA_{6a} index with three properties of octanes as well as GA_4 index with two properties of octanes. In order to further compare the descriptive and predictive abilities of these descriptors, we constructed for each of the properties of octanes a single regression model using only invariants with $|r| > 0.8$. Thus, Table 3 contains the values of s and Q^2 of five equations of the general form $BP = a + bGA$.

Table 3. Statistical parameters of equation $BP = a + bGA$ for five geometric-arithmetic indices[1].

	GA_1	GA_{6d}	GA_{6f} (α=0.835)	GA_{7b} (α=0.005)	$\mathbf{GA_8}$
s	3.581	2.684	3.469	6.117	**2.551**
Q^2	0.539	0.713	0.588	-0.121	**0.759**

[1] The best model is in bold.

In this case, the model based on GA_8 index outperforms all other models. In the case of the model based on GA_8, the improvement in the statistical deviation is equal to 28.76 % compared to the model based on GA_1 index. Note that while the linear correlation between GA_{7b} index at α=0.005 and the values of BPs is is greater than 0.8, the regression model based on this descriptor is devoid of any predictive capabilities. Table 4 presents the statistical parameters of ten equations of the general form $S = a + bGA$.

Table 4. Statistical parameters of equation $S = a + bGA$ for ten geometric-arithmetic indices[1].

	GA_1	GA_4	GA_{6a}	GA_{6b}	GA_{6c}	$\mathbf{GA_{6e}}$ **(β=0.555)**	GA_{6f} (α=0.5875)	GA_{7a} (β=1.93)	GA_{7b} (α=0.93)	GA_9
s	1.915	2.771	1.792	1.389	2.160	**1.266**	1.327	1.307	1.476	1.967
Q^2	0.756	0.516	0.793	0.866	0.679	**0.895**	0.880	0.882	0.845	0.735

[1] The best model is in bold.

In this case, the model based on GA_{6e} index at β=0.555 surpasses all other models. The improvement in the standard deviation is equal to 33.89 % relative to the model based on the first geometric-arithmetic index. All models exhibit good predictive abilities. In the case of the dataset of octane isomers, the density is satisfactorily linearly correlated only with GA_{6c} index. The regression model of the form $DENS = a + bGA_{6c}$ has the values of s and Q^2 equal to 0.014 and 0.368, respectively. Table 5 includes the statistical metrics of eight regression equations of the general form $HVAP = a + bGA$. In this case, the model based on GA_8 index possesses the best statistical characteristics.

Table 5. Statistical parameters of equation $HVAP = a + bGA$ for eight geometric-arithmetic indices[1].

	GA_1	GA_{6b}	GA_{6d}	GA_{6e} (β=0.005)	GA_{6f} (α=0.0025)	GA_{7a} (β=0.555)	GA_{7b} (α=0.005)	$\boldsymbol{GA_8}$
s	0.704	1.025	0.686	0.884	0.729	0.876	2.026	**0.670**
Q^2	0.802	0.680	0.831	0.716	0.821	0.711	-0.121	**0.855**

[1] The best model is in bold.

The improvement in the standard deviation is equal to 4.83 % relative to the model based on GA_1 index. The model based on GA_{7b} descriptor at α=0.005 has unsatisfactory predictive abilities. The standard enthalpy of vaporization is satisfactorily linearly correlated with nine geometric-arithmetic indices. The values of s and Q^2 of nine regression equations of the general form $DHVAP = a + bGA$ are reported in Table 6:

Table 6. Statistical parameters of equation $DHVAP = a + bGA$ for nine geometric-arithmetic indices[1].

	$\boldsymbol{GA_1}$	GA_{6b}	GA_{6d}	GA_{6e} (β=0.055)	GA_{6f} (α=0.075)	GA_{7a} (β=0.88)	GA_{7b} (α=0.51)	GA_8	GA_9
s	**0.103**	0.149	0.137	0.136	0.117	0.131	0.114	0.144	0.218
Q^2	**0.895**	0.819	0.845	0.823	0.879	0.833	0.891	0.831	0.629

[1] The best model is in bold.

In this case, the model based on GA_1 index has the best statistical parameters. In the case of this model, the improvement in the standard deviation is equal to 52.75 % versus the model based on GA_9 index (the worth statistical parameters). Table 7 presents the values of s and Q^2 of seven regression equations of the general form $HFORM = a + bGA$.

Table 7. Statistical parameters of equation $HFORM = a + bGA$ for seven geometric-arithmetic indices[1].

	GA_1	GA_{6d}	GA_{6e} (β=0.005)	GA_{6f} (α=0.0025)	GA_{7a} (β=0.005)	GA_{7b} (α=0.005)	$\boldsymbol{GA_8}$
s	0.661	0.5043	0.688	0.590	1.251	1.251	**0.418**
Q^2	0.685	0.806	0.653	0.752	-0.121	-0.121	**0.867**

[1] The best model is in bold.

In this case, the model based on GA_8 index is superior to all other models. The improvement in the standard deviation is equal to 36.76 % in comparison with the model based on the GA_1 index. Two models (i.e., the model based on GA_{7a} index at β=0.005 and the model based on GA_{7b} index at α=0.005) lack any predictive abilities. The accentric factor is satisfactorily linearly correlated with ten geometric-arithmetic indices.

Table 8. Statistical parameters of equation $ACENFAC = a + bGA$ for ten geometric-arithmetic indices[1].

	GA_1	GA_4	GA_{6a}	GA_{6b}	GA_{6c}	$\boldsymbol{GA_{6e}}$ **(β=1.005)**	GA_{6f} (α=0.805)	GA_{7a} (β=2.605)	GA_{7b} (α=0.97)	GA_9
s	0.015	0.018	0.007	0.006	0.019	**0.0032**	0.0033	0.0037	0.0046	0.008
Q^2	0.798	0.698	0.955	0.970	0.345	**0.990**	0.990	0.986	0.979	0.933

[1] The best model is in bold.

The statistical parameters of ten regression equations of the general form $ACENFAC = a + bGA$ corresponding to these descriptors are listed in Table 8. In this case, the best statistical parameters are exhibited by the model based on GA_{6e} index at β=1.005. The improvement in the standard deviation is equal to 78.67 %. All models possess very good predictive abilities. The motor octane number is linearly correlated with $|r| > 0.8$ with eight geometric-arithmetic indices. The values of s and Q^2 of eight regression equations of the general form $MON = a + bGA$ are presented in Table 9.

Table 9. Statistical parameters of equation $MON = a + bGA$ for 8 geometric-arithmetic indices[1].

	GA_{6a}	GA_{6b}	GA_{6c}	$\boldsymbol{GA_{6e}}$ **(β=1.355)**	GA_{6f} (α=0.865)	GA_{7a} (β=3.28)	GA_{7b} (α=0.9825)	GA_9
s	10.91	9.924	14.92	**8.541**	8.594	9.533	8.884	9.904
Q^2	0.751	0.803	0.556	**0.846**	0.846	0.810	0.837	0.802

[1] The best model is in bold.

The best statistical parameters are possessed by the model based on GA_{6e} index at β=1.355. The improvement in the value of s is equal to 42.75 % compared to the model based on the GA_{6c} index (the worth statistical metrics). All models have good predictive capabilities. The molar volume is satisfactorily linearly correlated with only one geometric-arithmetic index, i.e., with GA_{6c} descriptor. The regression equation of the form $MV = a + bGA_{6c}$ has the values of s and Q^2 equal to 2.872 and 0.42, respectively.

From Tables 2-9, it can be inferred that in many cases the regression models based on the new geometric-arithmetic indices perform considerably better than the regression models based on

the first geometric-arithmetic descriptor. In the case of the dataset of octane isomers, GA_1 index does not exhibit any satisfactory linear correlations with such properties as the density, the motor octane number and the molar volume. On the other hand, these properties are linearly correlated with $|r| > 0.8$ with GA_{6c} invariant (DENS and MV) or with GA_{6a}, GA_{6b}, GA_{6c}, GA_{6e} (at β=1.355), GA_{6f} (at α=0.865), GA_{7a} (at β=3.28) and GA_{7b} (at β=0.9825) indices (MON). In the case of such properties of octanes as the boiling point, the entropy, the enthalpy of vaporization, the enthalpy of formation and the acentric factor regression models based on one of the newly proposed geometric-arithmetic descriptors exhibit the improvement in the standard deviation from 4.83 % (HVAP) to 78.67 % (ACENFAC) compared to models based on the first geometric-arithmetic index.

The Pearson correlation coefficients between geometric-arithmetic indices (whose formulae do not include the adjustable parameters α or β) defined on the dataset of octane isomers are listed in Table 10.

The lowest linear correlation is noted between GA_8 and GA_4 indices (0.493) while the highest linear correlation is observed between GA_8 and GA_{6d} indices (0.995).

Table 10. Correlation coefficients between eight geometric-arithmetic indices defined on the dataset of octane isomers.

GA_1	1							
GA_4	0.700	1						
GA_{6a}	0.829	0.890	1					
GA_{6b}	0.961	0.822	0.948	1				
GA_{6c}	0.786	0.805	0.848	0.831	1			
GA_{6d}	0.960	0.531	0.651	0.848	0.690	1		
GA_8	0.949	0.493	0.613	0.827	0.640	0.995	1	
GA_9	0.852	0.839	0.976	0.958	0.767	0.688	0.662	1
	GA_1	GA_4	GA_{6a}	GA_{6b}	GA_{6c}	GA_{6d}	GA_8	GA_9

4.2 Correlations to the boiling points of saturated alkanes

Our initial studies have indicated that in the case of the boiling points of C_2-C_9 saturated alkanes, the polynomial regression produces better models that the single regression. Consequently, we obtained twelve equations of the general form $BP = a + b_1GA + b_2GA^2$ whose statistical parameters are presented in Table 11.

Table 11. Regression and statistical parameters of equation $BP = a + b_1 GA + b_2 GA^2$ for twelve geometric-arithmetic indices[1].

No	GA index	a	b_1	b_2	R^2	F	s	Q^2	$SDEP$
1b	GA_1	78.6559 (±0.5156)	300.7592 (±3.2198)	-33.7264 (±3.2198)	0.9959	4417.41	3.2198	0.9952	3.3728
2b	GA_4	78.6559 (±0.9948)	299.6998 (±6.2126)	-27.5201 (±6.2126)	0.9849	1173.381	6.2126	0.9816	6.5951
3b	GA_{6a}	78.6559 (±0.9858)	299.7704 (±6.1565)	-27.2038 (±6.1565)	0.9852	1195.204	6.1565	0.9825	6.4196
4b	GA_{6b}	78.6559 (±0.9844)	299.8800 (±6.1477)	-26.0422 (±6.1477)	0.9852	1198.667	6.1477	0.9826	6.4005
5b	GA_{6c}	83.0581 (±0.8167)	249.1928 (±5.0343)	-16.1779 (±5.0343)	0.9860	1230.23	5.0343	0.9837	5.2038
6b	$\boldsymbol{GA_{6d}}$	**78.6559 (±0.3921)**	**301.1991 (±2.4486)**	**-32.0946 (±2.4486)**	**0.9977**	**7651.634**	**2.4486**	**0.9966**	**2.8426**
7b	GA_{6e} (β=0.855)	78.6559 (±0.8518)	300.5076 (±5.3194)	-25.3641 (±5.3194)	0.9889	1607.098	5.3194	0.9870	5.5378
8b	GA_{6f} (α=0.9975)	78.6559 (±0.8801)	300.4528 (±5.4959)	-24.6489 (±5.4959)	0.9882	1504.37	5.4959	0.9862	5.7046
9b	GA_{7a} (β=1.405)	78.6559 (±0.6022)	301.0255 (±3.7610)	-29.0046 (±3.7610)	0.9945	3232.915	3.761	0.9936	3.874
10b	GA_{7b} (α=0.9425)	78.6559 (±0.6797)	300.5712 (±4.2446)	-31.2275 (±4.2446)	0.9929	2534.331	4.2446	0.9918	4.3913
11b	GA_8	78.6559 (±0.5910)	301.4876 (±3.6907)	-24.1181 (±3.6907)	0.9947	3357.778	3.6907	0.9934	3.9307
12b	GA_9	78.6559 (±0.8799)	300.4553 (±5.4952)	-24.6235 (±5.4952)	0.9882	1504.75	5.4952	0.9862	5.7038

[1] The best model is in bold.

With respect to the goodness of fit, the models from Table 11 can be ordered as follows:

Eq 6b (GA_{6d}) > Eq 1b (GA_1) > Eq 11b (GA_8) > Eq 9b (GA_{7a} (β=1.405)) > Eq 10b (GA_{7b} (α=0.9425)) > Eq 7b (GA_{6e} (β=0.855)) > Eq 12b (GA_9) > Eq 8b (GA_{6f} (α=0.9975)) > Eq 5b (GA_{6c}) > Eq 4b (GA_{6b}) > Eq 3b (GA_{6a}) > Eq 2b (GA_4)

The best statistical parameters are possessed by the model based on GA_{6d} index. The improvement in the statistical deviation is 23.95 % relative to the model based on GA_1 index. The results of t-test demonstrated that all variables in this model are significant. The model of

Eq 6b explains more than 99.76 % of the variance in the experimental values of BP for 39 alkanes. Table 12 shows the values of GA_{6d} index, the experimental boiling points of 39 saturated alkanes as well as the calculated (using Eq 6b) boiling points for this set of compounds.

Table 12. Experimental and calculated (with Eq 6b) boiling points (BPs) of 39 saturated alkanes with values of GA_{6d} index.

Subset	Compound	BP(°C)		GA_{6d}
		Exptl	Calcd	
A	ethane	-88.630	-84.956	1
C	propane	-42.070	-44.794	1.886
A	butane	-0.500	-1.609	2.931
A	2-methylpropane	-11.730	-14.856	2.598
B	pentane	36.075	35.150	3.922
C	2-methylbutane	27.852	28.568	3.736
C	2, 2-dimethylpropane	9.503	8.789	3.200
C	hexane	68.740	68.087	4.922
A	2-methylpentane	60.271	60.979	4.695
C	3-methylpentane	63.282	65.769	4.847
B	2, 3-dimethylbutane	57.988	57.628	4.590
A	2, 2-dimethylbutane	49.741	51.300	4.396
A	heptane	98.427	96.858	5.922
B	2-methylhexane	90.052	90.819	5.699
A	3-methylhexane	91.850	93.886	5.811
B	3-ethylpentane	93.475	96.824	5.920
B	2, 4-dimethylpentane	80.500	83.796	5.450
A	2, 2-dimethylpentane	79.197	80.452	5.334
C	2, 3-dimethylpentane	89.784	90.415	5.685
A	3, 3-dimethylpentane	86.064	87.456	5.579
C	2, 2, 3-trimethylbutane	80.882	78.889	5.281
A	octane	125.655	121.462	6.922
B	2-methylheptane	117.647	116.350	6.699
A	3-methylheptane	118.925	119.053	6.816
B	4-methylheptane	117.709	118.128	6.776
B	2, 5-dimethylhexane	109.103	111.087	6.479
C	3-ethylhexane	118.534	120.772	6.891
B	2, 4-dimethylhexane	109.429	113.252	6.569
B	2, 2-dimethylhexane	106.840	107.687	6.341
B	2, 3-dimethylhexane	115.607	115.229	6.652
C	3, 4-dimethylhexane	117.725	118.295	6.783
C	3, 3-dimethylhexane	111.969	112.073	6.520
A	3-ethyl-2-methylpentane	115.650	117.202	6.736
C	2, 2, 4-trimethylpentane	99.238	100.951	6.077
B	2, 3, 4-trimethylpentane	113.467	111.904	6.513
A	3-ethyl-3-methylpentane	118.259	117.339	6.742
C	2, 2, 3-trimethylpentane	109.840	108.250	6.364
B	2, 3, 3-trimethylpentane	114.760	110.455	6.453
C	2, 2, 3, 3-tetramethylbutane	106.470	98.640	5.989

The results of the y-randomization (after 1000 repetitions) produced the average value of R^2_{yrand} equal to 0.0241 and the average value of Q^2_{yrand} equal to -0.0824. Therefore, the

model based on GA_{6d} index does not include chance correlations. The results of external validation of this model are presented in Table 13.

Table 13. Results of external validation of model based on GA_{6d} index.

Training set	Prediction set	s	R^2
BC	A	2.9788	0.9988
AC	B	2.6556	0.9917
AB	C	3.1727	0.9970
	Average	2.8690	0.9958

The high average value of R^2 and the relatively low average value of s indicate that the model of Eq 6b has good predictive abilities with respect to external data. The calculated BPs versus the experimental data are depicted in Figure 3.

From the statistical considerations and Figure 3, we can see that the model based on GA_{6d} index is quite excellent. Note that the model of Eq 6b exhibits a lower standard deviation than models based on Xu index ($s = 5.791$), the Randić index χ (s=7.908), the molecular topological index (abbreviated as MTI) (s=17.975) and on the Hosoya Z index (s=22.924) [35, 41].

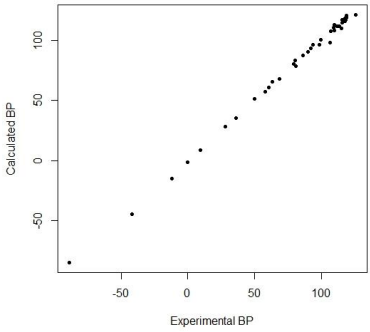

Figure 3. Plot of calculated boiling points (BP) of 39 alkanes versus experimental data.

4.3 Correlations to the enthalpies of combustion of aliphatic alcohols

Our preliminary studies have demonstrated that the enthalpies of combustion of aliphatic alcohols can be adequately modelled by the single regression. Thus, we obtained twelve

equation of the general form $\Delta_c H^\circ = a + bGA$. The statistical parameters of these models are detailed in Table 14.

Table 14. Regression and statistical parameters of equation $\Delta_c H^\circ = a + bGA$ for twelve geometric-arithmetic Indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1c	GA_1	-273.5408 (±35.7171)	-644.0825 (±3.6160)	0.9991	31726.22	110.431	0.9990	114.0153
2c	GA_4	-116.5654 (±9.1124)	-649.7894 (±0.9121)	>0.9999	507545.2	27.6208	>0.9999	29.481
3c	GA_{6a}	-105.5768 (±6.8934)	-650.5732 (±0.6898)	>0.9999	889382.5	20.8657	>0.9999	22.5035
4c	GA_{6b}	-90.6849 (±5.9390)	-651.6210 (±0.5942)	>0.9999	1202765	17.9428	>0.9999	19.402
5c	GA_{6c}	-188.6600 (±31.9697)	-644.5356 (±3.1483)	0.9994	41911.15	93.1386	0.9993	97.1657
6c	GA_{6d}	-237.8455 (±34.0321)	-644.7192 (±3.4330)	0.9992	35268.79	104.7426	0.9991	108.2022
7c	GA_{6e} (β=0.13)	-65.5072 (±5.2247)	-652.5242 (±0.5218)	>0.9999	1564092	15.7344	>0.9999	17.0682
8c	GA_{6f} (α=0.275)	-68.3668 (±5.2113)	-652.3988 (±0.5205)	>0.9999	1570980	15.6999	>0.9999	17.0682
9c	GA_{7a} (β=0.505)	-66.2138 (±5.2164)	-652.4761 (±0.5209)	>0.9999	1568796	15.7108	>0.9999	17.0606
10c	**GA_{7b} (α=0.605)**	**-69.2630 (±5.1988)**	**-652.2771 (±0.5192)**	**>0.9999**	**1578171**	**15.664**	**>0.9999**	**17.0784**
11c	GA_8	-126.3731 (±14.0846)	-652.3937 (±1.4172)	0.9999	211910.8	42.7445	0.9999	44.1835
12c	GA_9	-131.5371 (±10.8353)	-655.5411 (±1.0962)	>0.9999	357596.6	32.9057	>0.9999	34.0676

[1] The best model is in bold.

With respect to the goodness of fit, the models from Table 14 can be put in the following order:

Eq 10c (GA_{7b} (α=0.605)) > Eq 8c (GA_{6f} (α=0.275)) > Eq 9c (GA_{7a} (β=0.505)) > Eq 7c (GA_{6e} (β=0.13)) > Eq 4c (GA_{6b}) > Eq 3c (GA_{6a}) > Eq 2c (GA_4) > Eq 12c (GA_9) > Eq 11c (GA_8) > Eq 5c (GA_{6c}) > Eq 6c (GA_{6d}) > Eq 1c (GA_1).

The best results are obtained by the model based on GA_{7b} index at α=0.605. The relationship between the coefficient of determination (R^2) and the adjustable parameter α (of GA_{7b} descriptor) is shown in Figure 4.

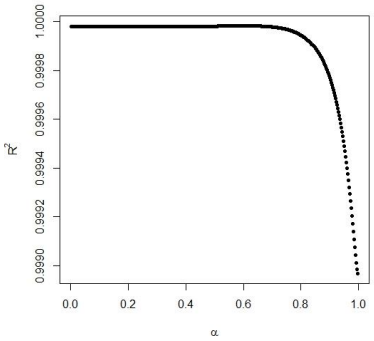

Figure 4. Plot of coefficient of determination (R^2) of equation 10c versus adjustable parameter α.

In the case of the model of Eq 10c, the improvement in the standard deviation is 85.82 % relative to the model based on the first geometric-arithmetic index. This model explains more than 99.99 % of the variance in the experimental data of $\Delta_c H^\circ$ for 29 aliphatic alcohols. Table 15 contains the values of GA_{7b} index at α=0.605, the experimental enthalpies of combustion of 29 aliphatic alcohols as well as the calculated (with Eq 10c) enthalpies of combustion for this set of compounds.

Table 15. Experimental and calculated (with Eq 7b) the enthalpies of combustion ($\Delta_c H^\circ$) of 29 aliphatic alcohols with values of GA_{7b} index at α=0.605.

Subset	Compound	$\Delta_c H^\circ$ (kJ/mol)		GA_{7b} (α=0.605)
		Exptl	Calcd	
B	methanol	-725.7	-721.54	1
A	ethanol	-1367.6	-1367.18	1.990
A	1-propanol	-2019.4	-2020.98	2.992
C	2-propanol	-2006.9	-2007.11	2.971
A	1-butanol	-2677.4	-2674.51	3.994
B	2-butanol	-2660.6	-2664.52	3.979
A	2-methyl-1-propanol	-2669.6	-2664.52	3.979
C	2-methyl-2-propanol	-2644	-2645.14	3.949
A	1-pentanol	-3324.6	-3327.46	4.995
C	2-pentanol	-3315.4	-3318.91	4.982
C	3-pentanol	-3312.3	-3320.25	4.984
C	2-methyl-1-butanol	-3325.9	-3320.25	4.984
B	3-methyl-1-butanol	-3326.2	-3318.91	4.982
B	2-methyl-2-butanol	-3303.1	-3303.43	4.958
B	3-methyl-2-butanol	-3315.1	-3313.85	4.974

B	1-hexanol	-3982.6	-3980.12	5.996
C	1-heptanol	-4642.52	-4632.62	6.996
A	1-octanol	-5295.5	-5285.05	7.996
B	1-nonanol	-5940.8	-5937.43	8.996
A	1-decanol	-6599.63	-6589.78	9.997
C	1-undecanol	-7253.7	-7242.12	10.997
B	1-dodecanol	-7909.4	-7894.44	11.997
B	1-tridecanol	-8517.8	-8546.75	12.997
C	1-tetradecanol	-9167	-9199.05	13.997
B	1-pentadecanol	-9817.7	-9851.35	14.997
A	1-hexadecanol	-10468.9	-10503.65	15.997
A	1-octadecanol	-11820	-11808	17.997
A	1-eicosanol	-13130	-13112.81	19.997
C	1-docosanol	-14450	-14417.38	21.997

Table 16. Results of external validation of model based on GA_{7b} index at α=0.605.

Training set	Prediction set	s	R^2
BC	A	15.6270	>0.9999
AC	B	18.6632	>0.9999
AB	C	18.9087	>0.9999
	Average	17.7330	>0.9999

The average values of R^2_{yrand} and Q^2_{yrand} after 1000 repetitions of the y-randomization are equal to 0.0345 and -0.1166, respectively. Thus, the model of Eq 10c does not have chance correlations. The results of external validation of the model based on GA_{7b} index at α=0.605 are shown in Table 16. These values testify that the above model possesses satisfactory predictive capabilities with respect to external data. The calculated enthalpies of combustion of 29 aliphatic alcohols versus the experimental data are presented in Figure 5.

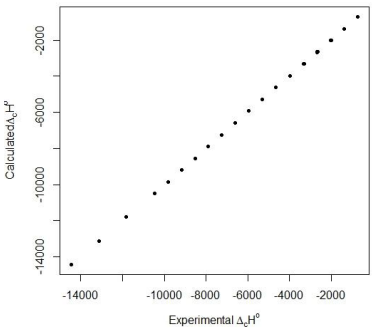

Figure 5. Plot of calculated enthalpies of combustion ($\Delta_c H^\circ$) of 29 aliphatic alcohols versus experimental data.

From the above facts, it can be deduced that the model of Eq 10c is very good.

4.4 Correlations to the molar volumes of alcohols

From our preliminary data, it can be seen that the molar volumes of aliphatic alcohols can be satisfactorily modelled by the single regression. Therefore, we obtained twelve linear equations of the general form $MV = a + bGA$ whose statistical characteristics are included in Table 17.

Table 17. Regression and statistical parameters of equation $MV = a + bGA$ for twelve geometric-arithmetic indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1d	GA_1	33.8438 (±2.1054)	16.2510 (±0.3209)	0.9846	2564.548	4.8806	0.9821	5.1362
2d	GA_4	28.1097 (±0.7179)	16.3053 (±0.1044)	0.9984	24376.14	1.5940	0.9980	1.7020
3d	GA_{6a}	27.7540 (±0.6842)	16.3631 (±0.0996)	0.9985	26996.75	1.5148	0.9982	1.6287
4d	GA_{6b}	27.1336 (±0.7045)	16.4368 (±0.1025)	0.9984	25731.94	1.5516	0.9981	1.6759
5d	GA_{6c}	30.8099 (±1.2233)	15.9997 (±0.1787)	0.9950	8016.497	2.7750	0.9943	2.9134
6d	GA_{6d}	32.4192 (±2.0855)	16.2830 (±0.3146)	0.9853	2679.673	4.7762	0.9829	5.0249
7d	GA_{6e} (β=0.005)	25.9763 (±0.6411)	16.4683 (±0.0925)	0.9987	31669.18	1.3988	0.9985	1.4997
8d	GA_{6f} (α=0.0025)	25.9755 (±0.6411)	16.4683 (±0.0925)	0.9987	31674.03	1.3987	0.9985	1.4996
9d	**GA_{7a} (β=0.005)**	**25.9755 (±0.6411)**	**16.4683 (±0.0925)**	**0.9987**	**31674.39**	**1.3987**	**0.9985**	**1.4996**
10d	**GA_{7b} (α=0.0025)**	**25.9755 (±0.6411)**	**16.4683 (±0.0925)**	**0.9987**	**31674.39**	**1.3987**	**0.9985**	**1.4996**
11d	GA_8	27.7760 (±1.4485)	16.6742 (±0.2149)	0.9934	6017.976	3.2002	0.9920	3.4372
12d	GA_9	28.0739 (±0.9590)	16.6984 (±0.1428)	0.9971	13665.28	2.1276	0.9963	2.3259

[1] The best models are in bold.

With respect to the descriptive properties, the models from Table 17 can be ordered as follows:

Eq 9d (GA_{7a} (β=0.005)) = Eq 10d (GA_{7b} (α=0.0025)) > Eq 8d (GA_{6f} (α=0.0025)) > Eq 7d (GA_{6e} (β=0.005)) > Eq 3d (GA_{6a}) > Eq 4d (GA_{6b}) > Eq 2d (GA_4) > Eq 12d (GA_9) > Eq 5d (GA_{6c}) > Eq 11d (GA_8) > Eq 6 (GA_{6d}) > Eq 1d (GA_1).

The best statistical parameters are possessed by the models based on GA_{7a} index at β=0.005 and on GA_{7b} index at α=0.0025. The relationship between the coefficient of determination (R^2) and the adjustable parameters β (of GA_{7a} invariant) or α (of GA_{7b} invariant) is presented in Figure 6.

Note that the values of GA_{7a} index at β=0.005 and GA_{7b} index at α=0.0025 are equal to the number of edges of a molecular graph. Consequently, it can be concluded that the molar volumes of 42 aliphatic alcohols are adequately modelled by a simple molecular descriptor, i.e., the number of edges of the corresponding H-depleted graph. In the cases of the models of Eqs 9d and 10d, the improvements in the standard deviation are equal to 71.34 % relative to the model based on GA_1 descriptor.

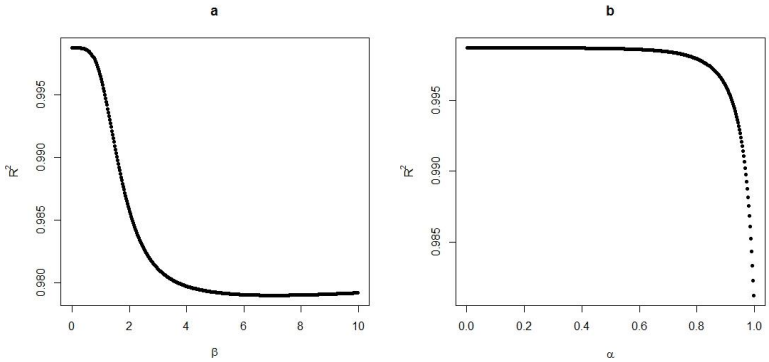

Figure 6. a/ Plot of coefficient of determination (R^2) of equation 9d versus adjustable parameter β, b/ Plot of coefficient of determination (R^2) of equation 10d versus adjustable parameter α.

These models account for more than 99.87 % of the variance in the experimental values of MVs of 42 alcohols. Table 20 presents the values of GA_{7a} index at β=0.005 (or GA_{7b} index at α=0.0025), the experimental molar volumes of 42 aliphatic alcohols as well as the calculated (with Eqs 9d or 10d) molar volumes for this set of compounds.

After 1000 repetitions, the y-scrambling produced the average values of R^2_{yrand} and Q^2_{yrand} equal to 0.0246 and -0.0762, respectively. Consequently, the above models do not have chance correlations. Table 18 presents the results of external validation of the models of Eqs 9d and 10d.

Table 18. Results of external validation of the model based on GA_{7a} index at β=0.005 (or on GA_{7b} index at α=0.0025).

Training set	Prediction set	s	R^2
BC	A	1.2300	0.9993
AC	B	1.5429	0.9990
AB	C	1.9416	0.9986
	Average	1.5715	0.9990

The high average value of R^2 and the low average value of s indicate that these models exhibit very good predictive abilities for external data. The plot of the calculated MVs of 42 aliphatic alcohols versus the experimental data is shown in Figure 7. It can be observed that the calculated values of MVs agree very well with the experimental data. Judging from the statistical parameters and plot in Figure 7, it can be uttered that the regression models based on GA_{7a} index at β=0.005 or on GA_{7b} index at α=0.0025 represent excellent QSPR models.

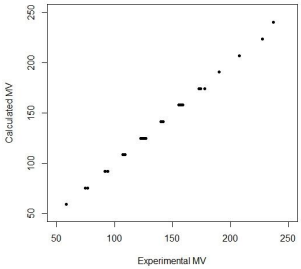

Figure 7. Plot of calculated molar volumes (MV) of 42 aliphatic alcohols versus experimental data.

For this same dataset, L. Mu et al. obtained the three-parameter regression model (with two edge connectivity indices, ie., ${}^0F$, ${}^1F$ and the alcohol-type parameter δ) with a slightly higher standard deviation (s=1.504) [37].

4.5 Correlations to the molar refractions of alcohols

The molar refraction (MR) is a measure of the total polarizability of molecules. This property is a particularly important physical characteristic in chemistry, biochemistry and pharmaceutical sciences. Our initial results have indicated that in the case of 41 aliphatic alcohols this property can be adequately described by the single regression. Therefore, we obtained twelve linear regression equations of the general form $MR = a + bGA$. The statistical parameters exhibited by these models are listed in Table 19.

Table 19. Regression and statistical parameters of equation $MR = a + bGA$ for twelve geometric-arithmetic indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1e	GA_1	5.7455 (±0.5141)	4.5620 (±0.0778)	0.9888	3437.633	1.1745	0.9871	1.2263
2e	GA_4	4.2815 (±0.0974)	4.5599 (±0.0141)	0.9996	104834.8	0.2138	0.9996	0.2244
3e	GA_{6a}	4.1675 (±0.0861)	4.5774 (±0.0125)	0.9997	135036.8	0.1884	0.9997	0.1984
4e	GA_{6b}	3.9960 (±0.0881)	4.5979 (±0.0127)	0.9997	130301.7	0.1918	0.9996	0.2041
5e	GA_{6c}	5.0053 (±0.2649)	4.4782 (±0.0385)	0.9971	13564.76	0.5937	0.9967	0.6201
6e	GA_{6d}	5.3520 (±0.4969)	4.5705 (±0.0744)	0.9898	3769.446	1.1221	0.9882	1.1728
7e	GA_{6e} (β=0.105)	3.7479 (±0.0815)	4.6044 (±0.0117)	0.9997	154562.6	0.1761	0.9997	0.1849
8e	**GA_{6f} (α=0.235)**	**3.7732 (±0.0814)**	**4.6031 (±0.0117)**	**0.9997**	**154622.6**	**0.1761**	**0.9997**	**0.1851**
9e	GA_{7a} (β=0.455)	3.7516 (±0.0816)	4.6043 (±0.0117)	0.9997	154037	0.1764	0.9997	0.1854
10e	GA_{7b} (α=0.5875)	3.7963 (±0.0818)	4.5997 (±0.0118)	0.9997	152840.8	0.1771	0.9997	0.1858
11e	GA_8	4.1199 (±0.3068)	4.6715 (±0.0452)	0.9964	10661.6	0.6694	0.9956	0.7138
12e	GA_9	4.2377 (±0.1789)	4.6736 (±0.0265)	0.9987	31131.43	0.3922	0.9985	0.4221

[1] The best model is in bold.

With regard to the goodness of fit, the models from Table 19 can be ordered as follows:

Eq 8e (GA_{6f} (α=0.235)) > Eq 7e (GA_{6e} (β=0.105)) > Eq 9e (GA_{7a} (β=0.455)) > Eq 10e (GA_{7b} (α=0.5875)) > Eq 3e (GA_{6a}) > Eq 4e (GA_{6b}) > Eq 2e (GA_4) > Eq 12e (GA_9) > Eq 5e (GA_{6c}) > Eq 11e (GA_8) > Eq 6e (GA_{6d}) > Eq 1e (GA_1).

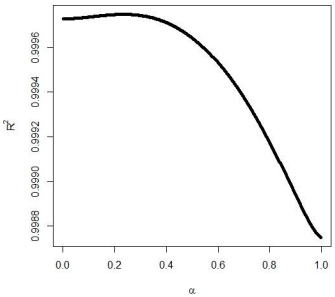

Figure 8. Plot of coefficient of determination (R^2) of equation 8e versus adjustable parameter α.

The model of Eq 8e shows the best statistical parameters. The relationship between the coefficient of determination (R^2) and the adjustable parameter α (of GA_{6f} invariant) is plotted in Figure 8. In the case of this model, the improvement in the statistical deviation is equal to 85.01 % relative to the model of Eq 1e. The model based on GA_{6f} index at α=0.235 is responsible for more than 99.97 % of the variance in the experimental MRs of 41 aliphatic alcohols. The values of GA_{6f} index at α=0.235, the experimental values of MRs as well as the calculated (with Eq 8e) values of MRs for this set of compounds are presented in Table 20.

Table 20. Experimental and calculated (with Eq 9d/10d (MV) or Eq 8e (MR)) the molar volumes (MVs) and the molar refractions (MRs) of aliphatic alcohols with values of GA_{7a} index at β=0.005 (or GA_{7b} index at α=0.0025) and GA_{6f} index at α=0.235.

Subset	Compound	MV(cm^3/mol)		GA_{7a} (β=0.005) GA_{7b} (α=0.0025)	MR(cm^3/mol)		GA_{6f} (α=0.235)
		Exptl	Calcld		Exptl	Calcd	
A	ethanol	58.368	58.912	2	12.927	12.959	1.996
C	1-propanol	74.798	75.380	3	17.565	17.561	2.995
B	2-propanol	76.561	75.380	3	17.613	17.504	2.983
C	1-butanol	91.529	91.849	4	22.145	22.166	3.996
B	2-methyl-1-propanol	92.338	91.849	4	22.182	22.117	3.985
C	2-butanol	91.903	91.849	4	22.144	22.117	3.985
B	2-methyl-2-propanol	94.216	91.849	4	22.033	22.014	3.963
A	1-pentanol	108.160	108.317	5	26.798	26.771	4.996
B	3-methyl-1-butanol	108.559	108.317	5	26.770	26.725	4.986
B	2-pentanol	108.962	108.317	5	26.724	26.725	4.986
B	2-methyl-1-butanol	108.027	108.317	5	26.753	26.730	4.987
A	3-pentanol	107.265	108.317	5	26.565	26.730	4.987
A	3-methyl-2-butanol	107.631	108.317	5	26.638	26.687	4.978
B	2-methyl-2-butanol	108.962	108.317	5	26.718	26.633	4.966
B	2, 2-dimethyl-1-propanol	108.559	108.317	5	-	-	-
C	1-hexanol	125.590	124.785	6	31.636	31.375	5.996
A	2-methyl-1-pentanol	123.795	124.785	6	31.262	31.337	5.988
C	2-ethyl-1-butanol	122.401	124.785	6	31.130	31.344	5.990
C	4-methyl-2-pentanol	126.774	124.785	6	31.497	31.292	5.978
C	2, 3-dimethyl-2-butanol	124.065	124.785	6	31.239	31.213	5.961
B	3, 3-dimethyl-1-butanol	124.005	124.785	6	31.224	31.242	5.968
B	3, 3-dimethyl-2-butanol	124.838	124.785	6	31.268	31.213	5.961
A	3-hexanol	124.716	124.785	6	31.297	31.337	5.988
A	3-methyl-3-pentanol	123.391	124.785	6	31.134	31.251	5.969
A	1-heptanol	141.345	141.253	7	36.015	35.978	6.996
C	2-heptanol	142.176	141.253	7	36.077	35.934	6.987
B	3-heptanol	141.535	141.253	7	35.981	35.942	6.988
A	4-heptanol	142.002	141.253	7	35.928	35.944	6.989
A	2, 4-dimethyl-3-pentanol	140.101	141.253	7	35.794	35.875	6.974
C	1-octanol	157.473	157.722	8	40.679	40.582	7.996
A	2-octanol	158.720	157.722	8	40.668	40.538	7.987
C	4-octanol	158.972	157.722	8	40.649	40.548	7.989
A	2-ethyl-1-hexanol	156.357	157.722	8	40.514	40.554	7.990
C	2, 2, 4-trimethyl-1-pentanol	155.221	157.722	8	40.097	40.432	7.964
A	3, 5-dimethyl-1-hexanol	156.960	157.722	8	40.135	40.510	7.981
B	1-nonanol	174.417	174.190	9	45.266	45.185	8.997
B	2, 6-dimethyl-4-heptanol	177.638	174.190	9	45.244	45.078	8.973

A	5-nonanol	172.642	174.190	9	44.589	45.152	8.989
B	1-decanol	190.252	190.658	10	49.734	49.788	9.997
C	1-undecanol	207.652	207.127	11	54.640	54.392	10.997
C	2, 6, 8-trimethyl-4-nonanol	227.438	223.595	12	59.289	58.862	11.968
C	1-tridecanol	236.965	240.063	13	63.375	63.598	11.997

In the case of the model of Eq 8e, the average values of R^2_{yrand} and Q^2_{yrand} after 1000 repetitions of the y-scrambling are equal to 0.0253 and -0.0792, respectively. Hence, the model of Eq 8e does not contain chance correlations. The results of external validation of the model based on GA_{6f} index at α=0.235 are shown in Table 21.

Table 21. Results of external validation of model based on GA_{6f} index at α=0.235.

Training set	Prediction set	s	R^2
BC	A	0.2452	0.9996
AC	B	0.0992	>0.9999
AB	C	0.2588	0.9998
	Average	0.2011	0.9998

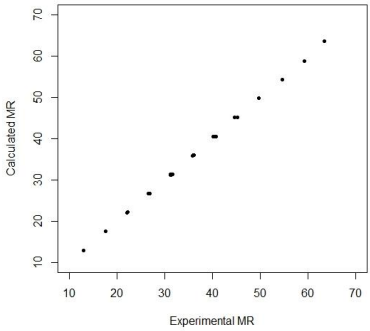

Figure 9. Plot of calculated molar refractions (MR) of 41 aliphatic alcohols versus experimental data.

The values of R^2 and s suggest that the above model exhibits satisfactory predictive capabilities with respect to external data. From the plot in Figure 9, it can be seen that the calculated values of MRs of 41 alcohols are very close to the experimental data. To sum up, the model based on GA_{6f} index at α=0.235 can be referred as very good. For this same dataset, L. Mu et al. obtained the three-parameter model (with ^0F, ^1F and δ as variables) with a higher standard deviation (s=0.446) [37].

4.6 Correlations to the molar refractions of aldehydes and ketones

Also, the molar refractions of the set of compounds composed of 22 aldehydes and 24 ketones are properly modelled by the single regression. So, twelve linear regression equations of the form $MR = a + bGA$ with their statistical parameters are presented in Table 22.

Table 22. Regression and statistical parameters of equation $MR = a + bGA$ for twelve geometric-arithmetic indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1f	GA_1	3.7920 (±0.3721)	4.6280 (±0.0515)	0.9946	8072.622	0.8642	0.9942	0.8780
2f	GA_4	2.7239 (±0.0950)	4.5891 (±0.0127)	0.9997	131361.6	0.2148	0.9996	0.2222
3f	GA_{6a}	2.6377 (±0.0848)	4.6017 (±0.0113)	0.9997	165811.4	0.1912	0.9997	0.1962
4f	GA_{6b}	2.4783 (±0.0817)	4.6217 (±0.0109)	0.9998	179960.2	0.1835	0.9997	0.1877
5f	GA_{6c}	3.0954 (±0.1501)	4.5479 (±0.0200)	0.9991	51605.61	0.3426	0.9990	0.3569
6f	GA_{6d}	3.4270 (±0.3650)	4.6299 (±0.0500)	0.9949	8559.956	0.8394	0.9945	0.8515
7f	GA_{6e} (β=0.13)	2.2712 (±0.0817)	4.6272 (±0.0108)	0.9998	182187.6	0.1824	0.9997	0.1869
8f	**GA_{6f} (α=0.315)**	**2.3176 (±0.0808)**	**4.6262 (±0.0107)**	**0.9998**	**186003.6**	**0.1805**	**0.9997**	**0.1849**
9f	GA_{7a} (β=0.48)	2.2635 (±0.0822)	4.6272 (±0.0109)	0.9998	180232.9	0.1834	0.9997	0.1879
10f	GA_{7b} (α=0.64)	2.3344 (±0.0815)	4.6226 (±0.0108)	0.9998	182320.5	0.1823	0.9997	0.1869
11f	GA_8	2.4545 (±0.2317)	4.7090 (±0.0315)	0.9980	22419.53	0.5195	0.9979	0.5264
12f	GA_9	2.6683 (±0.1527)	4.7015 (±0.0208)	0.9991	51000.95	0.3446	0.9991	0.3515

[1] The best model is in bold.

With respect to the decreasing goodness of fit, the models from Table 22 can be put in the following order:

Eq 8f (GA_{6f} (α=0.315)) > Eq 10f (GA_{7b} (α=0.64)) > Eq 7f (GA_{6e} (β=0.13)) > Eq 9f (GA_{7a} (β=0.48)) > Eq 4f (GA_{6b}) > Eq 3f (GA_{6a}) > Eq 2f (GA_4) > Eq 5f (GA_{6c}) > Eq 12f (GA_9) > Eq 11f (GA_8) > Eq 6f (GA_{6d}) > Eq 1 (GA_1).

Also, in this case, the best parameters are possessed by the model based on GA_{6f} index at α=0.315. Figure 10 presents the mutual relation between the coefficient of determination (R^2) and the adjustable parameter α (of GA_{6f} descriptor).

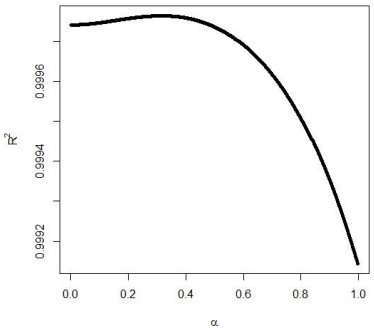

Figure 10. Plot of coefficient of determination (R^2) of equation 8f versus adjustable parameter α.

In the case of the model of Eq 8f, the improvement in the standard deviation is equal to 79.11 % compared to the model based on GA_1 index. This model elucidates more than 99.97 % of variance in the experimental data of MRs for this set of compounds. Table 23 presents the values of GA_{6f} index at α=0.315, the experimental values of the molar refractions of 22 aldehydes and 24 ketones as well as the calculated (with Eq 8f) values of MRs for this set of compounds.

Table 23. Experimental and calculated (with Eq 8f) the molar refractions of 22 aldehydes and 24 ketones with values of GA_{6f} index at α=0.315.

Subset	Compound	MR(cm³/mol)		GA_{6f} (α=0.315)
		Exptl	Calcd	
C	acetaldehyde	11.5829	11.5377	1.993
B	propionaldehyde	16.1632	16.1597	2.992
A	butyl aldehyde	20.8011	20.7878	3.993
A	2-methyl propanal	20.8219	20.7092	3.976
B	pentaldehyde	25.4983	25.4161	4.993
B	2-methyl butanal	25.3943	25.3473	4.978
A	3-methyl butanal	25.5327	25.3431	4.977
B	hexanal	30.0928	30.0438	5.993

B	2-methylpentanal	29.8497	29.9803	5.980
B	2-ethylbutanal	29.9981	29.9870	5.981
A	2, 3-dimethylbutanal	30.0640	29.9172	5.966
B	heptanal	34.7004	34.6710	6.994
A	2, 2-dimethylpentanal	34.7537	34.4780	6.952
C	octanal	39.4396	39.2979	7.994
C	2-ethylhexanal	39.2395	39.2480	7.983
C	2-ethyl-3-methylpentanal	38.9423	39.1985	7.972
C	nonanal	44.2669	43.9246	8.994
A	3, 5, 5-trimethylhexanal	43.9887	43.6785	8.941
B	decanal	48.6737	48.5512	9.994
A	2-methyldecanal	53.0003	53.1163	10.981
C	dodecanal	58.0913	57.8041	11.994
C	2-methylundecanal	57.9284	57.7426	11.981
B	acetone	16.2963	16.0722	2.973
C	2-butanone	20.6039	20.7092	3.976
A	2-pentanone	25.2926	25.3431	4.977
B	3-pentanone	25.2487	25.3473	4.978
A	3-methyl-2-butanone	25.2603	25.2765	4.963
B	2-hexanone	29.9308	29.9728	5.978
A	3-hexanone	29.7251	29.9803	5.980
C	3-methyl-2-pentanone	29.9453	29.9172	5.966
C	4-methyl-2-pentanone	29.9877	29.9100	5.964
A	3, 3-dimethyl-2-butanone	29.6748	29.7793	5.936
B	2-heptanone	34.5463	34.6008	6.978
C	3-heptanone	34.4230	34.6092	6.980
B	4-heptanone	34.3083	34.6126	6.981
A	5-methyl-2-hexanone	34.5773	34.5360	6.964
A	2-octanone	39.1959	39.2280	7.979
C	4-octanone	39.0616	39.2410	7.981
C	6-methyl-3-heptanone	38.9478	39.1724	7.967
A	2-nonanone	43.3542	43.8549	8.979
C	5-nonanone	43.8710	43.8692	8.982
A	2, 6-dimethyl-4-heptanone	43.8902	43.7491	8.956
B	2-decanone	48.5304	48.4816	9.979
C	2-undecanone	52.7129	53.1082	10.979
C	6-undecanone	53.2109	53.1230	10.982
B	2-methyl-4-undecanone	57.7027	57.6878	11.969

In the case of the model of Eq 8f, the y-randomization (after 1000 repetitions) gave the average value of R^2_{yrand} equal to 0.0229 and the average value of Q^2_{yrand} equal to -0.0703. The results of external validation of this model are presented in Table 24:

Table 24. Results of external validation of model based on GA_{6f} index at α=0.315.

Training set	Prediction set	s	R^2
BC	A	0.2189	0.9995
AC	B	0.1257	0.9999
AB	C	0.2158	0.9998
	Average	0.1868	0.9997

The values from Table 24 indicate that the model based on GA_{6f} index at α=0.315 has good predictive ability for external data. The plot of the calculated values of MRs of 22 aldehydes and 24 ketones versus the experimental data is depicted in Figure 11. This figure as well as all statistical conditions support the view that the model of Eq 8f can be considered as excellent. For this same dataset, B. Ren obtained the six-parameter model (with the modified Xu index, i.e., Xu_u^m and five atom-type-based AI topological indices) with a slightly lower standard deviation (s=0.1598) [43].

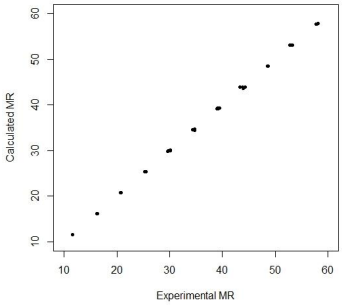

Figure 11. Plot of calculated molar refractions of 22 aldehydes and 24 ketones versus experimental data.

4.7 Correlations to the gas heat capacities of aldehydes and ketones

In the case of the gas heat capacities of the set of compounds composed of 3 aldehydes and 15 ketones, our preliminary studies have demonstrated that the power regression produces better models than any linear regression. Thus, we obtained twelve nonlinear equations of the general form $C_p^G = cGA^t$. The statistical metrics of these models are presented in Table 25.

Table 25. Regression and statistical parameters of equation $C_p^G = cGA^t$ for twelve geometric-arithmetic indices[1].

No	GA index	c	t	R^2	F	s	Q^2	$SDEP$
1g	GA_1	exp(3.6482) (±0.0302)	0.8214 (±0.0181)	0.9923	2051.123	0.0236	0.9903	0.0249
2g	GA_4	exp(3.5678) (±0.0421)	0.8376 (±0.0244)	0.9866	1179.175	0.0310	0.9837	0.0323
3g	GA_{6a}	exp(3.5527) (±0.0421)	0.8465 (±0.0244)	0.9869	1204.717	0.0306	0.9833	0.0326

4g	GA_{6b}	exp(3.5378) (±0.0437)	0.8540 (±0.0253)	0.9862	1141.826	0.0315	0.9823	0.0336
5g	GA_{6c}	exp(3.6362) (±0.0352)	0.8033 (±0.0205)	0.9897	1536.597	0.0272	0.9867	0.0290
6g	GA_{6d}	exp(3.6345) (±0.0345)	0.8221 (±0.0205)	0.9901	1603.255	0.0266	0.9877	0.0280
7g	GA_{6e} (β=1.505)	exp(3.5484) (±0.0355)	0.8588 (±0.0208)	0.9907	1702.669	0.0258	0.9878	0.0279
8g	GA_{6f} (α=0.9975)	exp(3.5420) (±0.0364)	0.8645 (±0.0214)	0.9903	1637.899	0.0263	0.9874	0.0284
9g	GA_{7a} (β=31.205)	exp(3.6338) (±0.0258)	0.8617 (±0.0161)	0.9945	2873.417	0.0199	0.9924	0.0220
10g	**GA_{7b} (α=0.97)**	**exp(3.6410) (±0.0255)**	**0.8291 (±0.0154)**	**0.9945**	**2915.997**	**0.0198**	**0.9927**	**0.0216**
11g	GA_8	exp(3.5387) (±0.0385)	0.8643 (±0.0226)	0.9892	1465.186	0.0278	0.9863	0.0295
12g	GA_9	exp(3.5418) (±0.0364)	0.8647 (±0.0214)	0.9903	1639.367	0.0263	0.9874	0.0284

[1] The best model is in bold.

With respect to the descriptive properties, the models from Table 25 can be ordered as follows:

Eq 10g (GA_{7b} (α=0.97)) > Eq 9g (GA_{7a} (β=31.205)) > Eq 1g (GA_1) > Eq 7g (GA_{6e} (β=1.505)) > Eq 12g (GA_9) > Eq 8g (GA_{6f} (α=0.9975)) > Eq 6g (GA_{6d}) > Eq 5g (GA_{6c}) > Eq 11g (GA_8) > Eq 3g (GA_{6a}) > Eq 2g (GA_4) > Eq 4g (GA_{6b}).

The model based on GA_{7b} index at α=0.97 outperforms all other models. The relationship between the coefficient of determination (R^2) and the adjustable parameter α (of GA_{7b} descriptor) is presented in Figure 12.

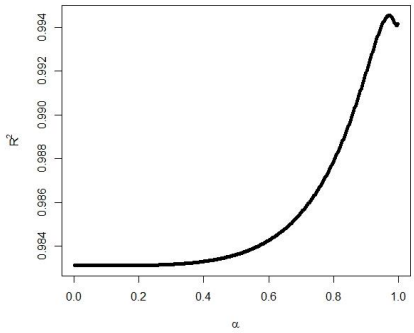

Figure 12. Plot of coefficient of determination (R^2) of equation 10g versus adjustable parameter α.

In the case of the model based on GA_{7b} index at α=0.97, the improvement in the standard deviation is equal to 16.10 % versus the model of Eq 1g. More than 99.45 % of the variance in the experimental data of the gas heat capacities of the considered set of compounds is explained by this model. The values of GA_{7b} index at α=0.97, the experimental gas heat capacities of the set of 3 aldehydes and 15 ketones as well as the calculated (with Eq 10g) values of C_p^G for this set of compounds are listed in Table 26.

Table 26. Experimental and calculated (with Eq 10g) gas heat capacities (C_p^G) of 3 aldehydes and 15 ketones with values of GA_{7b} index at α=0.97.

Subset	Compound	C_p^G(J/molK)		GA_{7b} (α=0.97)
		Exptl	Calcd	
B	propanal	90.03	90.64	2.842
A	pentanal	144.07	140.21	4.809
A	2, 2-dimethylpropanal	132.42	127.62	4.293
B	acetone	83.99	86.04	2.668
C	2-butanone	110.02	110.64	3.614
A	2-pentanone	136.23	135.57	4.618
C	3-pentanone	133.54	134.35	4.568
B	3-methyl-2-butanone	131.09	130.98	4.430
C	2-hexanone	161.50	159.99	5.639
A	3-hexanone	157.82	158.61	5.581
B	4-methyl-2-pentanone	155.68	156.34	5.484
C	3, 3-dimethyl-2-butanone	149.64	148.12	5.139
A	2-heptanone	189.55	183.69	6.662
C	4-heptanone	180.63	182.21	6.597
B	2-methyl-3-hexanone	173.25	178.28	6.426
B	2, 4-dimethyl-3-pentanone	171.98	174.70	6.270
A	2-octanone	209.54	206.70	7.680
C	5-nonanone	221.35	228.05	8.648

Table 27. Results of external validation of model based on GA_{7b} index at α=0.97.

Training set	Prediction set	s	R^2
BC	A	0.0388	0.9925
AC	B	0.0333	0.9987
AB	C	0.0222	0.9975
	Average	0.0314	0.9962

In the case of the model based on GA_{7b} index at α=0.97, the y-scrambling (after 1000 repetitions) produced the average value of R^2_{yrand} equal to 0.0562 and the average value of Q^2_{yrand} equal to -0.2065. Thus, this model does not possess chance correlations.

Table 27 presents the results of external validation of the model based on GA_{7b} index at α=0.97. The high average value of R^2 and the low value of s testify that this model can be reckoned as having very good predictive abilities with regard to external data. From the plot in Figure 13, it can be deduced that there exists agreement between the calculated (with Eq 10g) gas heat capacities and the experimental data. All aforementioned facts indicate that the model based on GA_{7b} index at α=0.97 is of high quality.

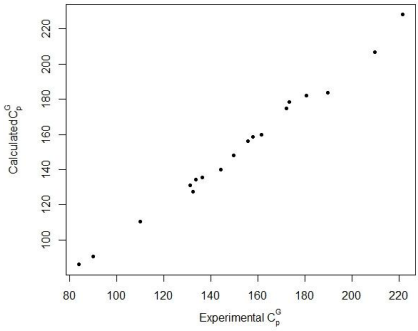

Figure 13. Plot of calculated gas heat capacities (C_p^G) of 3 aldehydes and 15 ketones versus experimental data.

For this same dataset, B. Ren obtained the two-parameter model (with Xu_u^m index and one AI index) with a higher standard deviation (s=2.48) [43].

4.8 Correlations to the enthalpies of formation of monocarboxylic acids

Our introductory findings suggest that the enthalpies of formation of 20 monocarboxylic acids can be satisfactorily modelled by the single regression. Hence, we obtained twelve linear regression equations of the general form $\Delta_f H^\circ = a + bGA$ whose statistical characteristics are presented in Table 28. With regard to the decreasing goodness of fit, the models from Table 28 can be arranged as follows:

Eq 11h (GA_8) > Eq 7h (GA_{6e} (β=9.98)) > Eq 9h (GA_{7a} (β=0.005)) = Eq 10h (GA_{7b} (α=0.0025)) > Eq 8h (GA_{6f} (α=0.0025)) > Eq 12h (GA_9) > Eq 4h (GA_{6b}) > Eq 3h (GA_{6a}) > Eq 2h (GA_4) > Eq 6h (GA_{6d}) > Eq 1h (GA_1) > Eq 5h (GA_{6c}).

Table 28. Regression and statistical parameters of equation $\Delta_f H^\circ = a + bGA$ for twelve geometric-arithmetic Indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1h	GA_1	-385.6264 (±4.9250)	-30.3508 (±0.3922)	0.9970	5989.068	10.0878	0.9962	10.8284
2h	GA_4	-379.1959 (±4.8123)	-30.1233 (±0.3747)	0.9972	6463.31	9.7118	0.9965	10.2969
3h	GA_{6a}	-378.328 (±4.7790)	-30.181 (±0.3720)	0.9973	6581.389	9.6245	0.9966	10.2119
4h	GA_{6b}	-377.7639 (±4.7006)	-30.2505 (±0.3663)	0.9974	6818.995	9.4558	0.9967	10.0279
5h	GA_{6c}	-381.1407 (±5.3052)	-29.9850 (±0.4131)	0.9966	5269.862	10.752	0.9957	11.5199
6h	GA_{6d}	-384.6374 (±4.9118)	-30.3259 (±0.3899)	0.997	6049.15	10.0378	0.9962	10.7668
7h	GA_{6e} (β=9.98)	-379.9932 (±4.6067)	-30.4219 (±0.3629)	0.9974	7027.406	9.3149	0.9968	9.9111
8h	GA_{6f} (α=0.0025)	-376.2199 (±4.6547)	-30.2804 (±0.3618)	0.9974	7003.77	9.3305	0.9968	9.8741
9h	GA_{7a} (β=0.005)	-376.2198 (±4.6547)	-30.2804 (±0.3618)	0.9974	7003.78	9.3305	0.9968	9.8741
10h	GA_{7b} (α=0.0025)	-376.2198 (±4.6547)	-30.2804 (±0.3618)	0.9974	7003.78	9.3305	0.9968	9.8741
11h	**GA_8**	**-379.3122 (±4.5586)**	**-30.6458 (±0.3612)**	**0.9975**	**7199.21**	**9.2033**	**0.9969**	**9.799**
12h	GA_9	-379.7984 (±4.6184)	-30.4641 (±0.3642)	0.9974	6998.026	9.3343	0.9968	9.9293

[1] The best model is in bold.

The model based on GA_8 index has the best descriptive properties. In the case of this model, the improvement in the standard deviation is equal to 8.77 % versus the model of Eq 1h. This model elucidates more than 99.75 % of the variance in the experimental values of $\Delta_f H^\circ$ of 20 monocarboxylic acids. Table 31 presents the values of GA_8 index, the experimental enthalpies of formation of 20 monocarboxylic acids as well as the calculated (with Eq 11h) values of $\Delta_f H^\circ$ for this set of compounds. In the case of the model of Eq 11h, the y-randomization (after 1000 repetitions) produced the average value of R^2_{yrand} equal to 0.0551 and the average value of Q^2_{yrand} equal to -0.1734. Thus, it can be claimed that the above model is devoid of any chance correlations. The results of external validation of the model based on GA_8 descriptor are listed in Table 29. These values indicate that the above model has a very good ability to predict external data.

The correlation between the calculated enthalpies of formation of 20 monocarboxylic acids and the experimental data is presented in Figure 14. In summary, it can be said that the model of Eq 8h is first-class.

Table 29. Results of external validation of the model based on GA_8 index.

Training set	Prediction set	s	R^2
BC	A	11.8822	0.9985
AC	B	14.6203	0.9960
AB	C	9.2483	0.9985
	Average	11.9169	0.9977

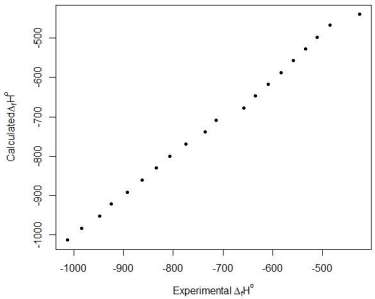

Figure 14. Plot of calculated enthalpies of formation ($\Delta_f H^\circ$) of 20 monocarboxylic acids versus experimental data.

The enthalpies of formation of monocarboxylic acids were modelled using the Randić and Harary indices by F. Shafiei [46]. But the dataset from [46] does not include formic acid. Also, this reference does not report the values of s for the obtained models. Nevertheless, it can be established here that for the dataset from Table 31, the single regression for $\Delta_f H^\circ$ produces the models with s=14.223 and R^2=0.9940 (using H index) and with s=10.3252 and R^2=0.9969 (using χ index). If we discard from this dataset formic acid, the single regression for $\Delta_f H^\circ$ gives the models with s=8.6373 and $R^2 = 0.9967$ (using GA_8 index), with s=9.9256 and R^2=0.9968[2] (using H index) and with s=9.3424 and R^2=0.9971[3] (using χ index). Consequently, it can be seen that regardless of whether the dataset contains formic acid or not, the models based on GA_8 index have lower values of the standard deviation.

4.9 Correlations to the enthalpies of combustion of monocarboxylic acids

Our initial studies have revealed that the enthalpies of combustion of 20 monocarboxylic acids can be adequately modelled by the single regression. Consequently, we obtained twelve linear regression equation of the general form $\Delta_c H^\circ = a + bGA$. The statistical parameters of these models are included in Table 30. With respect to the goodness of fit, the models from Table 30 can be put in the following order:

Eq 3i (GA_{6a}) > Eq 10i (GA_{7b} (α=0.745)) > Eq 4i (GA_{6b}) > Eq 8i (GA_{6f} (α=0.39)) > Eq 9i (GA_{7a} (β=0.63)) > Eq 7i (GA_{6e} (β=0.205)) > Eq 2i (GA_4) > Eq 12i (GA_9) > Eq 11i (GA_8) > Eq 6i (GA_{6d}) > Eq 1i (GA_1) > Eq 5i (GA_{6c}).

The best results are obtained by the model based on GA_{6a} index. In this case, the improvement in the standard deviation is equal to 78.98 % compared to the model of Eq 1i. This model explains more than 99.99 % of the variance in the experimental enthalpies of combustion of 20 monocarboxylic acids.

Table 30. Regression and statistical parameters of equation $\Delta_c H^\circ = a + bGA$ for twelve geometric-arithmetic Indices[1].

No	GA index	a	b	R^2	F	s	Q^2	$SDEP$
1i	GA_1	855.2810 (±15.581)	-650.691 (±1.241)	>0.9999	275039.1	31.9143	>0.9999	36.2354
2i	GA_4	992.5140 (±7.912)	-645.758 (±0.616)	>0.9999	1098804	15.9673	>0.9999	17.6785

[2] R^2=0.9971 [46].
[3] R^2=0.9981 [46].

3i	GA_{6a}	**1010.9791 (±3.3308)**	**-646.9847 (±0.2593)**	**>0.9999**	**6224582**	**6.7087**	**>0.9999**	**6.9075**
4i	GA_{6b}	1022.7190 (±3.9770)	-648.442 (±0.310)	>0.9999	4376603	8.007	>0.9999	8.2433
5i	GA_{6c}	952.770 (±20.919)	-642.963 (±1.629)	0.9999	155842	42.3964	0.9998	47.3893
6i	GA_{6d}	876.412 (±14.4000)	-650.150 (±1.143)	>0.9999	323477.2	29.4281	>0.9999	33.3352
7i	GA_{6e} (β=0.205)	1040.3737 (±4.3003)	-649.2262 (±0.3349)	>0.9999	3757957	8.6341	>0.9999	8.9478
8i	GA_{6f} (α=0.39)	1036.9111 (±4.1456)	-649.1207 (±0.3229)	>0.9999	4040626	8.3267	>0.9999	8.6069
9i	GA_{7a} (β=0.63)	1038.7054 (±4.2572)	-649.2316 (±0.3316)	>0.9999	3833134	8.5491	>0.9999	8.8403
10i	GA_{7b} (α=0.745)	1020.7715 (±3.4205)	-648.5517 (±0.2667)	>0.9999	5914927	6.8821	>0.9999	7.0205
11i	GA_{8}	988.849 (±13.689)	-656.855 (±1.085)	>0.9999	366752.9	27.6375	>0.9999	29.8392
12i	GA_{9}	978.7565 (±10.9781)	-652.9905 (±0.8656)	>0.9999	569048.8	22.1878	>0.9999	24.1843

[1] The best model is in bold.

The values of GA_{6a} invariant, the experimental enthalpies of combustion of 20 monocarboxylic acids as well as the calculated (with Eq 3i) values of $\Delta_c H^\circ$ for this set of compounds are listed in Table 31. In the case of the model of Eq 3i, the y-scrambling (after 1000 repetitions) gave the average values of R^2_{yrand} and Q^2_{yrand} equal to 0.0536 and -0.1768, respectively. The results of external validation of the model based on GA_{6a} index are contained in Table 32.

Table 31. Experimental and calculated (with Eq 11h or Eq 3i) the enthalpies of formation ($\Delta_f H^\circ$) and the enthalpies of combustion ($\Delta_c H^\circ$) of 20 monocarboxylic acids with values of GA_8 and GA_{6a} indices.

Subset	Compound	$\Delta_f H^\circ$ (kJ/mol)		GA_{10}	$\Delta_c H^\circ$ (kJ/mol)		GA_6
		Exptl	Calcd		Exptl	Calcd	
B	formic acid	-425.5	-439.98	1.979	-253.8	-256.85	1.960
A	ethanoic acid	-484.5	-467.42	2.875	-874.2	-868.34	2.905
C	propanoic acid	-510.8	-498.48	3.889	-1527.3	-1526.28	3.922
A	butanoic acid	-533.9	-527.75	4.844	-2183.5	-2184.75	4.939
B	pentanoic acid	-558.9	-557.37	5.810	-2837.8	-2839.69	5.952
B	hexanoic acid	-583.58	-587.26	6.786	-3492.4	-3492.07	6.960
A	heptanoic acid	-608.5	-617.35	7.767	-4146.9	-4142.86	7.966
C	octanoic acid	-634.8	-647.56	8.753	-4799.9	-4792.63	8.970

B	nonanoic acid	-658	-677.86	9.742	-5456.1	-5441.74	9.974
B	decanoic acid	-714.1	-708.23	10.733	-6079.3	-6090.40	10.976
C	undecanoic acid	-736.2	-738.65	11.726	-6736.5	-6738.74	11.978
B	dodecanoic acid	-775.1	-769.11	12.719	-7377	-7386.84	12.980
A	tridecanoic acid	-807.2	-799.60	13.714	-8024.2	-8034.76	13.981
B	tetradecanoic acid	-834.1	-830.11	14.710	-8676.7	-8682.55	14.983
A	pentadecanoic acid	-862.4	-860.63	15.706	-9327.7	-9330.24	15.984
A	hexadecanoic acid	-892.2	-891.18	16.703	-9977.2	-9977.83	16.985
C	heptadecanoic acid	-924.4	-921.73	17.700	-10624.4	-10625.35	17.985
A	octadecanoic acid	-948	-952.30	18.697	-11280.1	-11272.82	18.986
C	nonadecanoic acid	-984.1	-982.87	19.695	-11923.4	-11920.24	19.987
C	eicosanoic acid	-1012.6	-1013.45	20.693	-12574.2	-12567.61	20.987

The high average value of R^2 and the low average value of s testify that the above model is reliable with respect to external data. From the plot in Figure 15, it can be inferred that the calculated values (with Eq 3i) are in agreement with the experimental enthalpies of combustion of 20 monocarboxilic acids.

Table 32. Results of external validation of the model based on GA_{6a} index.

Training set	Prediction set	s	R^2
BC	A	6.8620	>0.9999
AC	B	10.3101	>0.9999
AB	C	7.0395	>0.9999
	Average	8.0705	>0.9999

Figure 15 and all statistical metrics suggest that the model of Eq 3i is very good. The enthalpies of combustion of monocarboxylic acids were modelled using the Randić and Harary indices by F. Shafiei [46]. But in [46], formic acid is excluded from studies. Also, this reference does not contain the values of the standard deviation for the models based on these invariants.

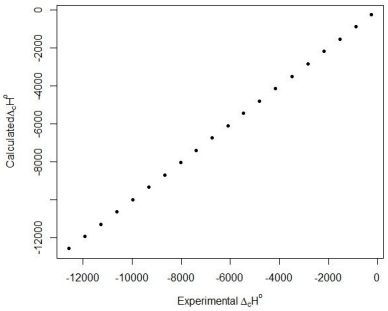

Figure 15. Plot of calculated enthalpies of combustion ($\Delta_c H^\circ$) of 20 monocarboxylic acids versus experimental data.

Nevertheless, it can be demonstrated here, that for the dataset from Table 31, the single regression for $\Delta_c H^\circ$ produces the model with s=248.075 and R^2=0.9946 (using H index) and with s=40.5239 and R^2=0.9999 (using χ index). If we remove from this dataset formic acid, the single regression for $\Delta_c H^\circ$ produces the models with s=6.8545 and R^2>0.9999 (using GA_{6a} index), with s=248.075 and R^2=0.9956[4] (using H index) and with s=10.5151 and R^2>0.9999[5] (using χ index). Therefore, it can be seen that regardless of if the dataset includes formic acid or not, the models based on GA_{6a} index have lower values of the standard deviation.

In the above paragraphs (b-i), we have presented eight QSPR models based on the newly defined topological indexes. All final models exhibited the values of R^2 and Q^2 above 0.99 and relatively low values of s. In all cases, the best statistical parameters were possessed by the model based on any of the newly defined molecular descriptors. Namely, the improvement in the standard deviation was in the range of 8.77 % (Eq 11h) to 85.82 % (Eq 10c) compared to the models based on the first geometric-arithmetic index. In four cases, the best models were based on the sixth geometric-arithmetic index (on its GA_{6d} (Eq 6b), GA_{6f} (at α=0.235, Eq 8e), GA_{6f} (at α=0.315, Eq 8f) and GA_{6a} (Eq 3i) versions). In three cases, the best models were based on the seventh geometric-arithmetic index (on its GA_{7b} (at α=0.605, Eq 10c), GA_{7a} (at β=0.005 Eq 9d) or GA_{7b} (at α=0.0025, Eq 10c) and GA_{7b} (at α=0.97 Eq 10g) versions). In one case, the best model was based on the eighth geometric-arithmetic index (Eq 11h). On the other hand, in four cases the worst models were based on the first geometric-arithmetic index (Eq 1c, Eq 1d, Eq 1e and Eq 1f). In two cases, the worst models were based on the fourth geometric-arithmetic index (Eq 2a and Eq 2g). Also in two cases, the worst models were based on the sixth geometric-arithmetic index (on its GA_{6c} (Eq 5h and Eq 5i) version). All final models explain more than 99 % of the variance in the experimental data. In all cases, the y-randomization (after 1000 repetitions) applied to the final models gave the average value of R^2 from 0.0229 (Eq 8f) to 0.0562 (Eq 10g) and the average value of Q^2 from -0.2065 (Eq 10g) to -0.0703 (Eq 8f). Therefore, it can be claimed that all final models are devoid of any chance correlations. In the procedure of external validation, all predictions were made with the value of R^2 above 0.99. Hence, all final models have very good predictive capabilities relative to external data.

[4] R^2=0.9934 [46].
[5] R^2=1 [46].

In conclusion, it can be stated that all the above presented models according to the criteria cited in Part Two are excellent and reliable with respect to external data.

5 Concluding remarks

In this work, we have introduced several new geometric-arithmetic indices. We have demonstrated that these newly defined descriptors have an extremely low level of degeneracy (Part Three) as well as in many cases exhibit better correlation properties than the first geometric-arithmetic index (Part Four).

It is hoped that the newly proposed molecular invariants will be widely used in QSAR/QSPR studies.

References

[1] N. Adriaanse, H. Dekker, J. Coops, Heats of combustion of normal saturated fatty acids and their methyl esters, *Recl. Trav. Chim. Pays-Bas* **84** (1965) 393-407.

[2] A. T. Balaban, O. Ivanciuc, Historical development of topological indices, in: J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999, pp. 21-57.

[3] M. Benzi, C. Klymko, Total communicability as a centrality measure, *J. Complex Networks* **1** (2013) 124-149.

[4] J. Chao, F. D. Rossini, Heats of combustion, formation and isomerization of nineteen alkanols, *J. Chem. Eng. Data* **10** (1965) 374-379.

[5] M. J. Crawley, *Statistics. An Introduction Using R*, Wiley, Chichester, 2015.

[6] J. J. Crofts, D. H. Higham, A weighted communicability measure applied to complex brain networks, *J. R. Soc. Interface* **6** (2009) 411-414.

[7] G. Csardi, T. Nepusz, The igraph software package for complex network research, *InterJournal* **Complex Systems 1695** (2006) http://igraph.org

[8] K. C. Das, On geometric-arithmetic index of graph, *MATCH Commun. Math. Comput. Chem.* **64** (2010) 619-630.

[9] K. C. Das, I. Gutman, B. Furtula, On second geometric-arithmetic index of graphs, *Iran. J. Math. Chem.* **1** (2010) 17-27.

[10] K. C. Das, I. Gutman, B. Furtula, On third geometric-arithmetic index of graphs, *Iran. J. Math. Chem.* **1** (2010) 29-36.

[11] K. C. Das, I. Gutman, B. Furtula, Survey on geometric-arithmetic indices of graphs, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 595-644.

[12] K. C. Das, I. Gutman, B. Furtula, On first geometric-arithmetic index of graphs, *Discr. Appl. Math.* **159** (2011) 2030-2037.

[13] K. C. Das, N. Trinajstić, comparison between geometric-arithmetic indices, *Croat. Chim. Acta* **85** (2012) 353-357.

[14] M. Dehmer, M. Grabner, B. Furtula, Structural discrimination of networks by using distance, degree and eigenvalue-based measures, *PLoS ONE* **7** (2012) 1-15.

[15] A. A. Dobrynin, A. A. Kocheova, Degree distance of a graph: a degree analog of the Wiener index, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1082-1086.

[16] E. Estrada, J. A. Rodriguez-Velazquez, Subgraph centrality in complex networks, *Phys. Rev. E* **71** (2005) #056103.

[17] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, *Proteomics* **6** (2006) 35-40.

[18] E. Estrada, Protein bipartivity and essentiality in the yeast protein-protein interaction network, *J. Proteome Res.* **5** (2006) 2177-2184.

[19] G. Fath-Tabar, B. Futula, I. Gutman, A new geometric-arithmetic index, *J. Math. Chem.* **47** (2010) 477-486.

[20] B. Freeman, M. O. Bagby, Heats of combustion of fatty esters and triglycerides, *J. Am. Oil Chem. Soc.* **66** (1989) 1601-1605.

[21] M. Ghorbani, A. Khaki, A note on the fourth version of geometric-arithmetic index, *Optoelectron. Adv. Mater. Rapid Commun.* **4** (2010) 2212-2215.

[22] A. Graovac, M. Ghorbani, M. A. Hosseinzadeh, Computing fifth geometric-arithmetic index for nanostar dendrimers, *J. Math. Nanosci.* **1** (2011) 33-42.

[23] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, *J. Comput. Graph. Stat.* **5** (1996) 299-314.

[24] O. Ivanciuc, Graph theory in chemistry, in: J. Gasteiger (Ed.), *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim, 2003, pp. 103-138.

[25] O. Ivanciuc, Topological indices, in: J. Gasteiger (Ed.), *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim, 2003, pp. 981-1003.

[26] O. Ivanciuc, A. T. Balaban, The graph description of chemical structures, in: J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999, pp. 59-167.

[27] D. Janežič, A. Miličević, S. Nikolić, N. Trinajstić, *Graph Theoretical Matrices in Chemistry*, Univ. Kragujevac, Kragujevac, 2007.

[28] B. S. Junkes, R. D. M. C. Amboni, R. A. Yunes, V. E. F. Heinzen, Prediction of the chromatographic retention of saturated alcohols on stationary phases of different polarity applying the novel semi-empirical topological index, *Anal. Chim. Acta* **477** (2003) 29-39.

[29] R. Kiralj, M. M. C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, *J. Braz. Chem. Soc.* **20** (2009) 770-787.

[30] C. F. Klymko, *Centrality and Communicability Measures in Complex Networks: Analysis and Algorithms*, PhD thesis, Emory University, Atlanta, GA 2013.

[31] E. V. Konstantinova, V. A. Skorobogatov, M. V. Vidyuk, Applications of information theory in chemical graph theory, *Indian J. Chem.* **42A** (2003) 1227-1240.

[32] N. D. Lebedeva, Heats of combustion of monocarboxylic acids, *Russ. J. Phys. Chem. (Engl. Transl.)* **38** (1964) 1435-1437.

[33] A. Mahmiani, O. Khormali, A. Iranmanesh, On the edge version of geometric-arithmetic index, *Dig. J. Nanomater. Biostruct.* **7** (2012) 411-414.

[34] A. Mahmiani, O. Khormali, On the total version of geometric-arithmetic index, *Iran. J. Math. Chem.* **4** (2013) 21-26.

[35] Z. Mihalić, N. Trinajstić, A graph-theoretical approach to structure-property relationships, *J. Chem. Educ.* **69** (1992) 701-712.

[36] C. Mosselman, H. Dekker, Enthalpies of formation of n-alkan-1-ols, *J. Chem. Soc., Faraday Trans.* **71** (1975) 417-424.

[37] L. Mu, C. Feng, L. Xu, Study on QSPR of alcohols with a novel edge connectivity index ^{m}F, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 217-230.

[38] NIST Chemistry WebBook, http://webbook.nist.gov/chemistry/

[39] A. Platzer, P. Perco, A. Lukas, B. Mayer, Characterization of protein-interaction networks in tumors, *BMC Bioinf.* **8** (2007) 224.

[40] G. Pólya, R. C. Read, *Combinatorial Enumeration of Groups, Graphs and Chemical Compounds*, Springer, New York, 1987.

[41] R Core Team (2015), R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[42] B. Ren, A new topological index for QSPR of alkanes, *J. Chem. Inf. Comput.* Sci. **39** (1999) 139-143.

[43] B. Ren, New atom-type-based AI topological indices: application to QSPR studies of aldehydes and ketones, *J. Comput. Aided Mol. Des.* **17** (2003) 607-620.

[44] J. M. Rodríguez, J. M. Sigarreta, On the geometric-arithmetic index, *MATCH Commun. Math. Comput. Chem.* **74** (2015) 103-120.

[45] J. M. Rodríguez, J. M. Sigarreta, Spectral study of the geometric-arithmetic index, *MATCH Commun. Math. Comput. Chem.* **74** (2015) 121-135.

[46] F. Shafiei, Relationship between topological indices and thermodynamic properties and of the monocarboxylic acids applications in QSPR, *Iran. J. Math. Chem.* **6** (2015) 15-28.

[47] H. A. Skinner, A. Snelson, The heats of combustion of the four isomeric butyl alcohols, *Trans. Faraday Soc.* **56** (1960) 1776-1783.

[48] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2009.

[49] D. Vukičević, B. Furtula, Topological index based on the ratios of geometrical and arithmetical means of end-vertex degrees of edges, *J. Math. Chem.* **46** (2009) 1369-1376.

[50] B. Zhou, I. Gutman, B. Furtula, Z. Du, On two types of geometric-arithmetic index, *Chem. Phys. Lett.* **482** (2009) 153-155.