OCWLGI Descriptors: Theory and Praxis

Andrey A. Toropov*,1, Alla P. Toropova1, Emilio Benfenati1 and Giuseppina Gini2

1IRCCS - Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy
2Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract: The aim of this review is description of the logic and evolution of optimal descriptors OCWLGI calculated with the molecular graph and the demonstration of their ability as tools for the modeling of biological and physicochemical parameters of chemical compounds. The ability of optimal descriptors calculated with hydrogen suppressed graph (HSG), hydrogen filled graph (HFG) and graph of atomic orbitals (GAO) is demonstrated as a collection of quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR) for properties and endpoints available from the literature. The Monte Carlo method optimization of the correlation weights of local and global invariants (OCWLGI) of molecular graphs is used as the principle for building up descriptors which are discussed in this article. The statistical quality of the QSPR and QSAR models for physicochemical and biological properties which were obtained with the optimal descriptors are reasonably high.

Keywords: Global invariant, local invariant, Monte Carlo method, optimization of correlation weights of local and global invariants of graph (OCWLGI), quantitative structure-property relationships, QSPR, quantitative structure-activity relationships, QSAR.

INTRODUCTION

Many quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR) models are influenced by the work of Wiener on molecular graphs, topological indices and prediction of physicochemical properties [1-4]. The main idea of these studies is the use of molecular graphs, molecular matrices and topological indices that may be correlated with biological and physicochemical properties of organic compounds. Starting in the 1980’s, the number of topological indices conceptually related to the Wiener number started to increase [5-14]. Most of these descriptors or indices were based on two matrices, namely the adjacency matrix and the distance matrix. Consider the molecular graph of 2-methyl butane, with the vertex labelling as shown in Fig. (1).

Fig. (1). The molecular graph of 2-methyl butane (CAS 78-78-4).

The adjacency matrix $A(G)$ and the distance matrix $D(G)$ of 2-methyl butane are:

<table>
<thead>
<tr>
<th>$A(G)$</th>
<th>$D(G)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>1 0 1 0 0</td>
<td>0 1 2 3 2</td>
</tr>
<tr>
<td>2 1 1 0 1</td>
<td>1 0 1 2 1</td>
</tr>
<tr>
<td>3 0 1 0 1</td>
<td>2 1 0 1 2</td>
</tr>
<tr>
<td>4 0 0 1 0</td>
<td>3 2 1 0 3</td>
</tr>
<tr>
<td>5 0 1 0 0</td>
<td>2 1 2 3 0</td>
</tr>
</tbody>
</table>

Although the topological indices that may be computed for chemical structures are diverse [15], the main principles in developing topological indices can be illustrated with the Wiener number ($W$) and connectivity indices of zero-order ($\chi^0$) and first-order ($\chi^1$), which is also known as the Randić index [8,12-15]. For example, these topological indices for 2-methyl butane (Fig. 1) are calculated with the formulae:

$$W = (1/2) \Sigma d_{ij} = (1/2)(0 + 1 + 2 + 3 + 2 + 1 + 0 + 1 + 2 + 1 + 0 + 1 + 2 + 3 + 0) = 18$$  \hspace{1cm} (1)

$$\chi^0 = \Sigma \delta_i^{1/2} = \delta_1^{1/2} + \delta_2^{1/2} + \delta_3^{1/2} + \delta_4^{1/2} = (1)^{1/2} + (3)^{1/2} + (2)^{1/2} + (1)^{1/2} = 1 + 0.577 + 0.707 + 1 + 1 = 4.284$$  \hspace{1cm} (2)

$$\chi^1 = \Sigma (\delta_i \delta_j)^{1/2} = (\delta_1 \delta_2)^{1/2} + (\delta_2 \delta_3)^{1/2} + (\delta_3 \delta_4)^{1/2} + (\delta_4 \delta_1)^{1/2} = (1 \cdot 3)^{1/2} + (3 \cdot 2)^{1/2} + (2 \cdot 1)^{1/2} + (3 \cdot 1)^{1/2} = 0.577 + 0.408 + 0.707 + 0.577 = 2.269$$  \hspace{1cm} (3)

where $\delta_i$ is the vertex degree equal to the sum of the elements of the adjacency matrix in a row, and $d_{ij}$ is the element of the $D(G)$ matrix. Detailed information on other structural descriptors calculated in the last years to develop QSPR and QSAR models may be found in [15, 16]. Some topics related to the investigation of the topological indices are related to their definition. For instance, why the exponent for the vertex degree in Eq. (2) and Eq. (3) have been selected as −1/2, and not another value? Indeed, the QSAR analysis of several molecular properties [17] has shown that in most cases the optimal value of the degree, in models based on descriptors similar to the Eq.(2) or Eq. (3), is different from −1/2.

QSPR and QSAR studies [18-21] have shown that the correction of the adjacency matrix (e.g. ethyl isopropyl
sulphide, CAS 5145-99-3, Fig. 2) by means of replacing the zero values on the diagonal of the adjacency matrix with special selected coefficients \(x\) and \(y\) (Fig. 3) may produce considerable improvement of correlation between the connectivity indices \(\chi^0\) and \(\chi^1\) and various properties.

![Graph](Image)

**Fig. (2).** The hydrogen-suppressed graph (HSG) of ethyl isopropyl sulphone (CAS 5145-99-3).

### ADJACENCY MATRIX

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & 0 & 1 & 0 & 0 & 0 \\
2 & 1 & 0 & 1 & 0 & 0 \\
3 & 0 & 1 & 0 & 1 & 0 \\
4 & 0 & 0 & 1 & 0 & 1 \\
5 & 0 & 0 & 0 & 1 & 0 \\
6 & 0 & 1 & 0 & 0 & 0 \\
\end{array}
\]

\[\delta_i\]

### MODIFIED ADJACENCY MATRIX

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & 5 & 6 \\
1 & x & 1 & 0 & 0 & 0 \\
2 & 1 & x & 1 & 0 & 0 \\
3 & 0 & 1 & y & 1 & 0 \\
4 & 0 & 0 & 1 & x & 1 \\
5 & 0 & 0 & 0 & 1 & x \\
6 & 0 & 1 & 0 & 0 & x \\
\end{array}
\]

\[\delta_i\]

It has to be noted that by adding these parameters on the diagonal of the adjacency matrix, the calculated values for \(\delta_i\) are also modified. Similar modifications have been carried out for the graph distance matrix [22], and the result was a significant improvement of the QSPR models. These descriptors have been named “variable” or “flexible”, however, we will use the term “optimal descriptors”, because descriptors which are discussed here are calculated by means of the Monte Carlo method optimization. All these optimal descriptors [18-22] have been computed from the hydrogen-suppressed molecular graph.

### CONSTRUCTION OF OPTIMAL DESCRIPTORS

The first optimal descriptor, based on the hydrogen-suppressed graph, has been suggested by Randić [23, 24]. The main idea of the approach was to use diagonal entries of the adjacency matrix to consider the influence of heteroatoms, similarly to the well known generalization of the Hückel molecular orbitals method [23].

The optimal descriptors have been successfully used for the QSPR modeling of aliphatic alcohols [25], nitrogen-containing compounds [26], and sulfides [27]. Later on, optimal descriptors based on the hydrogen-filled graph have been suggested [28].

The optimization target may be the standard error of estimation (\(s\)) [23-27] (i.e. searching for minimum of \(s\)), or the correlation coefficient (\(r\)) [28, 29] (i.e. searching for maximum of \(r\)), and by comparing the results obtained with these target functions it has been shown that the maximization of correlation coefficient gave models with better statistical quality [29].

### OPTIMAL DESCRIPTORS BASED ON HSG AND HFG

As an example of the general scheme based on the hydrogen-suppressed graphs we consider ethyl isopropyl sulphone (Fig. 2). Accordingly to [30], the use of \(x=+0.25\) and \(y=-0.95\) in the calculation of the optimal connectivity index \(\chi^0 (x,y)\) for the correlation with normal boiling points of 21 sulfides gave the more accurate model in comparison with the standard formula with fixed coefficients \(\chi^0 (0,0)\).

An optimal descriptor is a modification of fixed coefficients (e.g., vertex degree, path of length 2, etc.), in the calculation of a “classic” topological index. This modification is made by using parameters that improve the statistical indices of the QSPR and QSAR models. In other words, each descriptor is a mathematical function of the representation of the molecular structure (MSR). Any MSR contains molecular invariants (MI), which define the molecular individuality. An MSR can be represented by

\[D = F(MSR) = F(MI_1, MI_2, ..., MI_m)\]  (4)

where \(MI_k\) is the \(k\)-th molecular invariant \((k = 1, ..., m, \) where \(m\) is the total number of molecular invariants in the molecule). The descriptor that is calculated with Eq. (4) is the fixed version which is similar to descriptors calculated with Eq.(1), Eq.(2), and Eq.(3).

Formula (4) can be modified by replacing the fixed components \(MI_k\) with flexible ones \(CW(MI_k)\):

\[D = F(MSR) = F(CW(MI_1), CW(MI_2), ..., CW(MI_m))\]  (5)

where \(CW(MI_k)\) is the correlation weight of the \(k\)-th molecular invariant. The descriptor \(D\) calculated with Eq. (5) is a flexible version of the descriptor calculated with Eq. (4). The correlation weights \(CW(MI_k)\) are numerical coefficients used in the calculation with Eq.(5).

The correlation coefficient between a descriptor calculated with Eq. (5) and the property/activity (PA) of interest is also a mathematical function of the CWs,

\[R(PA, D) = R(PA, F(CW(MI_1), CW(MI_2), ..., CW(MI_m))\]  (6)

where \(R(PA, D)\) is the correlation coefficient between the PA and \(D\), calculated with Eq. (5).

The optimization of the parameters \(CW(MI_1), CW(MI_2), ..., CW(MI_m)\) is performed with the Monte Carlo method, which results in a maximum value for the \(R(PA, D)\) for the training set of compounds. The predictive ability of the model can be tested with an external set of compounds.

As a possible extension of the scheme based on the Eq. (6), these correlation weights can be calculated not only for...
The optimal descriptor based on the hydrogen-filled graph and the hydrogen-filled graph (HFG) based optimal phenomena (property/activity) under consideration. Be used as a hint on the study of mechanism related to the molecular graph) on the property/activity of interest, that can influence for a given molecular attribute (invariant of invariants in the modeling) one can estimate a measure of the descriptors can be named OCWLGI descriptors.

Owing to this possibility (involving of local and global invariants in the modeling) one can estimate a measure of the influence for a given molecular attribute (invariant of molecular graph) on the property/activity of interest, that can be used as a hint on the study of mechanism related to the phenomena (property/activity) under consideration.

A comparison of the hydrogen-suppressed graph based and the hydrogen-filled graph (HFG) based optimal descriptors has been carried out [32]. It has been shown that the optimal descriptor based on the hydrogen-filled graph gave better models for the normal boiling points of alkyl alcohols.

OCWLGI DESCRIPTORS BASED ON GAO

The graph of atomic orbitals (GAO) is an attempt to take into account the structure of atoms in QSPR/QSAR analysis [31-37]. The conversion of the hydrogen filled graph into GAO can be carried out by the scheme:

1. Each vertex of the hydrogen filled graph is replaced by the group of atomic orbitals. Such groups of the atomic orbitals are listed in Table 1.

2. Elements of the adjacency matrix of the graph of atomic orbitals are defined as

\[ a_{ij} = \begin{cases} 1, & \text{if } i\text{-th and } j\text{-th GAO vertices fall in different atoms in HFG and these atoms have joint edge in HFG} \\ 0, & \text{otherwise} \end{cases} \]

The groups of atomic orbitals for some chemical elements are listed in Table 1.

For a training set of graphs of atomic orbitals, one can carry out the same Monte Carlo optimization of correlation weights of the invariants using the same algorithms [31-38]. However, the models based on the HSG or HFG and GAO-based model are different. For the same list of compounds the number of different vertexes as well as vertex degree values for the GAO representation is larger. This approach improves the statistical quality of a model (calculated with the optimal descriptors) for the training set, but the statistical quality for the external test set can be poor [31]. Consequently, one should be careful with the GAO representation of chemical compounds, because it may lead to overtraining [31]. However, quite good GAO-based models are also possible [31-34].

For the hydrogen-filled graph of nitrapyrin (CAS 1929-82-4) with vertex numbering shown in Fig. (4), this conversion gives the GAO shown in Fig. (2).

![Fig. (3). The numbered HFG of nitrapyrin (CAS 1929-82-4).](image)

**Table 1. Groups of Atomic Orbitals, for Some Chemical Elements, Used in Constructing the Graph of Atomic Orbitals**

<table>
<thead>
<tr>
<th>Chemical Elements</th>
<th>Group of Atomic Orbitals</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>1s¹</td>
</tr>
<tr>
<td>C</td>
<td>1s², 2s², 2p²</td>
</tr>
<tr>
<td>N</td>
<td>1s², 2s², 2p³</td>
</tr>
<tr>
<td>O</td>
<td>1s², 2s², 2p³</td>
</tr>
<tr>
<td>F</td>
<td>1s², 2s², 2p⁵</td>
</tr>
<tr>
<td>P</td>
<td>1s², 2s², 2p⁶, 3s², 3p⁵</td>
</tr>
<tr>
<td>S</td>
<td>1s², 2s², 2p⁶, 3s², 3p⁵</td>
</tr>
<tr>
<td>Cl</td>
<td>1s², 2s², 2p⁶, 3s², 3p⁵</td>
</tr>
<tr>
<td>Br</td>
<td>1s², 2s², 2p⁶, 3s², 3p⁵</td>
</tr>
</tbody>
</table>

The QSPR/QSAR models in this study are characterized by the number of compounds in a dataset (\(n\)); the square of correlation coefficient (\(r^2\)); standard error of estimation (\(s\)); and Fischer F-ratio. The generalized form of the optimal descriptor is the following:

\[
DCW = CW(G) + \left\{ \sum_{i=1}^{N} CW(A_i) + \sum_{i=1}^{N} W(VI_k) \right\} \tag{7}
\]

where \(N\) is the number of vertex in molecular graph, i.e. the number of atoms in the case of HSG (Fig. 2) and HFG (Fig. 4) and the number of AO in the case of GAO (Fig. 5); \(CW(A_i)\) is the correlation weight of chemical element (or AO in the case of the GAO) which is an image of the k-th vertex; \(CW(VI_k)\) is the correlation weight of the invariant of k-th vertex such as vertex degree, Morgan extended connectivity [31, 38], valence shells [9, 47], the number of paths of length 2 or 3 [47]; \(CW(G)\) is the correlation weight of a global invariant of the molecular graph, such as the number of cycles [7, 47], hydrogen bond indices [39], etc. As an alternative to the additive scheme for optimal descriptor (i.e. Eq. 7), the multiplicative scheme [31] can be used. Table 2 contains a collection of QSPR/QSAR models which were built up with OCWLGI descriptors.
DISCUSSION

The model for normal boiling points [48] (developed with the same chemical compounds as for Model 7, Table 2) is characterized by $n = 134, r^2 = 0.9886, s = 2.86 \, ^\circ C, F = 3770$. Thus, the statistical characteristics for the Model 7 are better. The statistical indices of the best QSPR for molar heat capacity (at 300$^\circ$K) [48] are $n = 134, r^2 = 0.9771, s = 3.89 \, J/K-1 \, mol^{-1}, F = 1848$. The statistical characteristics of the Model 8 are better (Table 2). The best QSPR for standard Gibbs energy of formation in the gas phase at 300 $^\circ$K has $n = 134, r^2 = 0.9172, s = 4.29 \, kJ/mol, F = 480$ [48]. The statistical quality of the Model 9 is better. The best model for alkane vaporization enthalpy at 300$^\circ$K is characterized by $n = 134, r^2 = 0.9172, s = 4.29 \, kJ/mol, F = 480$ [48]. The statistical quality of the Model 9 is better. The best model for alkane vaporization enthalpy at 300$^\circ$K is characterized by $n = 134, r^2 = 0.9172, s = 4.29 \, kJ/mol, F = 480$ [48]. The statistical quality of the Model 9 is better. The best QSPR for refractive index $n_D^{25}$ [48] is characterized by $n = 134, r^2 = 0.9683, s = 0.0025, F = 1309$. The statistical quality of the Model 10 is better. Finally, the statistical quality of the best model for alkane density is $n = 134, r^2 = 0.9805, s = 3.73 \, kg/m^3, F = 2156$ [48]. The statistical quality of the Model 12 is better. It is to be noted, that in ref. [48] the external test set is absent.

A QSAR analysis based on the quantum chemical descriptors for 57 anti-HIV-1 agents of tetrahydroimidazole[4,5,l-\textit{jk}][1,4]-benzodiazepin-2-(1\textit{H})-one (TIBO) derivatives together with 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives is presented. The QSAR model for TIBO derivatives has $r^2 = 0.8649, s = 0.597$ [49], and the statistical quality of the QSAR model for HEPT derivatives is $r^2 = 0.9063, s = 0.371$. The Model 15 is related to both TIBO and HEPT derivatives. The model is checked with the external test dataset. Thus, the statistical quality of Model 15 and statistical quality of models from ref. [49] should be considered as similar.

The statistical indices of the best model for the acute aquatic toxicity are $n = 69, r^2 = 0.863, s = 0.30$ [50]. The statistical characteristics for Model 19 are better. In conclusion, these comparisons with models for the same endpoints from the literature indicate that the optimal OCWLGI-descriptors can be useful for the QSPR/QSAR analyses. Finally, we note that there are several topological indices which can be translated into their OCWLGI type of molecular descriptors [54-61].
Table 2. A Collection of QSPR/QSAR Based on the OCWLGI-Descriptors

<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Statistical Characteristics</th>
<th>Refs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physicochemical Parameters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>QSPR models for normal boiling points of alkanes, alkylbenzenes, and polyaromatic hydrocarbons</td>
<td>$n = 70$, $r^2 = 0.9988$, $s = 5.8^\circ C$, $F = 57437$ (training set) $n = 70$, $r^2 = 0.9985$, $s = 6.7^\circ C$, $F = 45154$ (test set)</td>
<td>[38]</td>
</tr>
<tr>
<td>2</td>
<td>QSPR modelling of the constants of stability of 110 biometal $M^{2+}$ complexes with $\alpha$-amino acids and phosphate derivatives of adenine</td>
<td>$n = 55$, $r^2 = 0.9843$, $s = 0.279$, $F = 3328$ (Training set) $n = 55$, $r^2 = 0.9935$, $s = 0.248$, $F = 4027$ (Test set)</td>
<td>[39]</td>
</tr>
<tr>
<td>3</td>
<td>QSPR model of normal boiling points of alcohols</td>
<td>$n = 29$, $r^2 = 0.9906$, $s = 9.2^\circ C$, $F = 5733$ (training set) $n = 29$, $r^2 = 0.9896$, $s = 3.0^\circ C$, $F = 2595$ (test set)</td>
<td>[32]</td>
</tr>
<tr>
<td>4</td>
<td>QSPR model of normal boiling points of acyclic carbonyl compounds</td>
<td>$n = 100$, $r^2 = 0.9972$, $s = 6.12^\circ C$, $F = 3464$ (training set) $n = 100$, $r^2 = 0.9975$, $s = 6.00^\circ C$, $F = 3905$ (test set)</td>
<td>[40]</td>
</tr>
<tr>
<td>5</td>
<td>QSPR model of normal boiling points of haloalkanes (GAO)</td>
<td>$n = 138$, $r^2 = 0.9841$, $s = 9.80^\circ C$, $F = 3905$ (training set) $n = 138$, $r^2 = 0.9854$, $s = 7.39^\circ C$, $F = 3905$ (test set)</td>
<td>[35]</td>
</tr>
<tr>
<td>6</td>
<td>The challenged study involved predictions of normal boiling points for organic molecules of varied composition. These molecules included species with both linear and cyclic structures, comprise ketones, esters, aldehydes, nitriles, amines, alcohols, and hydrocarbons and a wide variety of atoms, such as C, H, O, N, Si, Cl, Br, F, P, and S</td>
<td>$n = 126$, $r^2 = 0.9279$, $s = 33.3^\circ C$, $F = 1599$ (training set); $n = 32$, $r^2 = 0.8819$, $s = 39.1^\circ C$, $F = 224$ (test set)</td>
<td>[41]</td>
</tr>
<tr>
<td>7</td>
<td>The short list of additional examples relevant to this review includes: Normal boiling points of alkanes</td>
<td>$n = 67$, $r^2 = 0.9984$, $s = 1.126^\circ C$, $F = 39180$ (training set); $n = 67$, $r^2 = 0.9910$, $s = 2.553^\circ C$, $F = 7118$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>8</td>
<td>Molar heat capacity at 300 K ($J/(K mol)$) of alkanes</td>
<td>$n = 67$, $r^2 = 0.9892$, $s = 2.790$, $F = 5869$ (training set); $n = 67$, $r^2 = 0.9902$, $s = 2.442$, $F = 6067$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>9</td>
<td>Standard Gibbs energy of formation in the gas state at 300 K ($kJ/mol$) of alkanes</td>
<td>$n = 67$, $r^2 = 0.9984$, $s = 1.804$, $F = 5465$ (training set); $n = 67$, $r^2 = 0.9903$, $s = 1.791$, $F = 3224$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>10</td>
<td>Vaporization enthalpy at 300 K ($kJ/mol$) of alkanes</td>
<td>$n = 67$, $r^2 = 0.9735$, $s = 0.0024$, $F = 2385$ (training set); $n = 67$, $r^2 = 0.9627$, $s = 0.0024$, $F = 1684$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>11</td>
<td>Refractive index at 25°C of alkanes</td>
<td>$n = 67$, $r^2 = 0.9844$, $s = 3.602$, $F = 4036$ (training set) $n = 67$, $r^2 = 0.9763$, $s = 3.790$, $F = 2680$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>12</td>
<td>Density ($kg/m^3$) of alkanes</td>
<td>$n = 67$, $r^2 = 0.9844$, $s = 3.602$, $F = 4036$ (training set) $n = 67$, $r^2 = 0.9763$, $s = 3.790$, $F = 2680$ (test set)</td>
<td>[31]</td>
</tr>
<tr>
<td>13</td>
<td>Flory-Huggins parameter for binary polymer-solvent mixtures</td>
<td>$n = 30$, $r^2 = 0.9990$, $s = 0.028$, $F = 27537$ (training set); $n = 30$, $r^2 = 0.9972$, $s = 0.053$, $F = 10294$ (test set)</td>
<td>[42]</td>
</tr>
<tr>
<td>14</td>
<td>The intrinsic viscosity of polymers</td>
<td>$n = 17$, $r^2 = 0.9130$, $s = 0.126 cm^3/g$, $F = 157$ (training set); $n = 9$, $r^2 = 0.9231$, $s = 0.143 cm^3/g$, $F = 84$ (test set).</td>
<td>[42]</td>
</tr>
<tr>
<td>Biological Activity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Anti-HIV-1 activity TIBO and HEPT derivatives</td>
<td>$n = 37$, $r^2 = 0.8688$, $s = 0.557$, $F = 232$ (training set) $n = 20$, $r^2 = 0.8759$, $s = 0.588$, $F = 127$ (test set)</td>
<td>[43]</td>
</tr>
<tr>
<td>16</td>
<td>Toxicity, V. fischeri, log(1/IGC$_{50}$), valence shells has been used as local graph invariant</td>
<td>$n = 45$, $r^2 = 0.8299$, $s = 0.402$, $F = 210$ (training set) $n = 21$, $r^2 = 0.8902$, $s = 0.339$, $F = 154$ (test set)</td>
<td>[44]</td>
</tr>
<tr>
<td>17</td>
<td>Toxicity to Tetrahymena pyriformis of heterogeneous set of benzene derivatives</td>
<td>$n = 157$, $r^2 = 0.883$, $s = 0.27$, $F = 1170$ (training set); $n = 60$, $r^2 = 0.863$, $s = 0.28$, $F = 372$ (test set);</td>
<td>[45]</td>
</tr>
<tr>
<td>18</td>
<td>The mutagenic activities of 95 heteroaromatic compounds in S. typhimurium TA98 S9, graph of atomic orbitals</td>
<td>$n = 47$, $r^2 = 0.7637$, $s = 1.05$, $F = 145$ (training set); $n = 48$, $r^2 = 0.7569$, $s = 0.86$, $F = 144$ (test set)</td>
<td>[34]</td>
</tr>
<tr>
<td>19</td>
<td>Aquatic toxicity, Pimephales promelas, log(1/LC$_{50}$), Morgan extended connectivity is used as local graph invariant</td>
<td>$n = 233$, $r^2 = 0.7689$, $s = 0.75$, $F = 258$ (training set) $n = 41$, $r^2 = 0.6421$, $s = 1.14$, $F = 70$ (test set)</td>
<td>[47]</td>
</tr>
<tr>
<td>20</td>
<td>Acute toxicity LC$_{50}$-96h to rainbow trout (Oncorhynchus mykiss) of 274 organic pesticides</td>
<td>$n = 36$, $r^2 = 0.8857$, $s = 0.500$, $F = 279$ (training set) $n = 36$, $r^2 = 0.9251$, $s = 0.382$, $F = 414$ (test set)</td>
<td>[52]</td>
</tr>
<tr>
<td>21</td>
<td>Carcinogenic activity of methylated polycyclic aromatic hydrocarbons</td>
<td>$n=36$, $r^2=0.6446$, $s=0.861$, $F=62$ (training set) $n=37$, $r^2=0.7843$, $s=0.616$, $F=142$ (test set)</td>
<td>[53]</td>
</tr>
<tr>
<td>22</td>
<td>Lipophilicity (logP) of 76 industrial chemicals</td>
<td>$n=36$, $r^2=0.8909$, $s=0.689$ (training set) $n=16$, $r^2=0.9247$, $s=0.594$ (test set)</td>
<td>[51]</td>
</tr>
<tr>
<td>23</td>
<td>The mutagenic activities of these compounds in S. typhimurium TA100 + S9 microsomal preparation are expressed in log of revertant per nonamole, ln $R$.</td>
<td>$n=36$, $r^2=0.8909$, $s=0.689$ (training set) $n=16$, $r^2=0.9247$, $s=0.594$ (test set)</td>
<td>[51]</td>
</tr>
</tbody>
</table>
CONCLUSIONS

The OCWLGI descriptors calculated from the molecular graph are a robust approach for the predictive modeling of physicochemical and biomedical properties. Our experience with the OCWLGI descriptors may be summarized in the following rules to obtained high quality QSPR and QSAR models: (1) complex molecular descriptors, such as the Morgan extended connectivity, or valence shells, may result in overfitting and poor prediction statistics; (2) the QSPR and QSAR models based on the OCWLGI descriptors should be validated by splitting the dataset into a training dataset and a test dataset, and then repeating several times this procedure; (3) infrequent structural descriptors computed from the molecular graph should be removed from the process of training a QSPR or QSAR model. The CORAL software (http://www.insilico.eu/coral/) is available for building up described models.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The authors express their gratitude to CALEIDOS (the project number LIFE11-INV/IT 00295) for financial support. Also the authors express their gratitude to Dr. L. Cappellini, Dr. R. Bagnati, and Dr. G. Bianchi for technical assistance. This paper is dedicated to Professor Milan Randić on the occasion of his 80th birthday.

ABBREVIATIONS

| OCWLGI | = Optimization of correlation weights of local and global invariants of graph |
| QSPR/QSAR | = Quantitative structure-property/activity relationships |
| HFG | = Hydrogen-filled graph |
| HSG | = Hydrogen-suppressed graph |
| GAO | = Graph of atomic orbitals |

REFERENCES

[8] King, R.B. In: Chemical applications of topology and graph theory. Proceedings of a symposium held at the University of Georgia, Athens, Georgia, USA, 1983