ORIGINAL SCIENTIFIC PAPER

CROATICA CHEMICA ACTA

Croat. Chem. Acta **2025**, *98*(4), 267–278
Published online: December 31, 2025
DOI: 10.5562/cca4199

HKD 1926

Novel Laplacian Matrix-based Molecular Descriptors Derived by Graph Convolution: Development and Applications in QSAR Studies

Igor Kuzmanovski,[1,*] Subhash C. Basak[2]

[1] Institute of Chemistry, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University in Skopje, Macedonia

[2] Department of Chemistry and Biochemistry, University of Minnesota Duluth, USA

* Corresponding author's e-mail address: shigor@pmf.ukim.mk

RECEIVED: July 28, 2025 * REVISED: December 11, 2025 * ACCEPTED: December 12, 2025

Abstract: This article reports the development of a set of new molecular descriptors derived from convolution using the Laplacians of molecular graphs and their line graphs. These descriptors have been applied in quantitative structure–activity relationship (QSAR) studies to predict the toxicity of 69 benzene derivatives and the aqueous solubility of a diverse dataset of 375 drug-like structures, using multivariate linear regression and a nonlinear machine learning algorithm known as counter-propagation artificial neural networks. The descriptors are developed using atomic properties of only the non-hydrogen atoms in the molecule. Using this approach, we developed a total of 54 new graph convolution-based descriptors. Results indicate that the newly defined invariants provide a new set of molecular descriptors for the characterization of molecular structures and QSAR studies.

Keywords: molecular graph, line graph, adjacency matrix, laplacian matrix, laplacian-transformed property matrix, quantitative structure-activity relationship (QSAR), aquatic toxicity, benzene derivatives, aqueous solubility, machine learning method, graph convolution descriptors.

INTRODUCTION

A recent trend in the development of quantitative structure-activity relationships (QSARs) of molecules and biomolecules is the use of computed chemodescriptors and biodescriptors along with modeling tools derived from statistics[1–9] and machine learning approaches.[10–13] In many practical situations related to new drug discovery and toxicological evaluation of chemicals for the protection of human and environmental health, we need to know a set of important physicochemical and biological properties of many chemicals that are untested or not yet synthesized. The experimental determination of properties of such large number of candidate chemicals is prohibitively costly or impractical. A practical way to address such a quagmire is to use quantitative structure-activity relationship (QSAR) models based on molecular descriptors, which can be calculated directly from molecular structure without the input of any other experimental data. Such descriptors are mainly topological, three-dimensional (3-D) or quantum chemical in nature. Mathematical methods like matrix theory, graph theory, and information theory are some frequently used tools in the computation of molecular descriptors.[5–10] Use of descriptors in predictive QSAR models arises from the dictum: *Function (property) follows form (structure)*. These days often the number of predictors/descriptors (p) that can be calculated using various available software is much larger than n, the number of data points, to be modeled. In such rank-deficient cases, proper statistical and machine learning methods need to be used to develop robust predictive models. Therefore, it is evident that robust statistical/machine learning (ML) methods constitute an indispensable link between structural/property data on the one hand and implementable models in the end user's computer, on the other.

Beginning with the pioneering publications of Wiener[14] at the middle of the twentieth century, the field of development of graph invariants or graph theoretic molecular descriptors has witnessed a major growth spurt through the major contributions of Balaban,[15] Randić,[16]

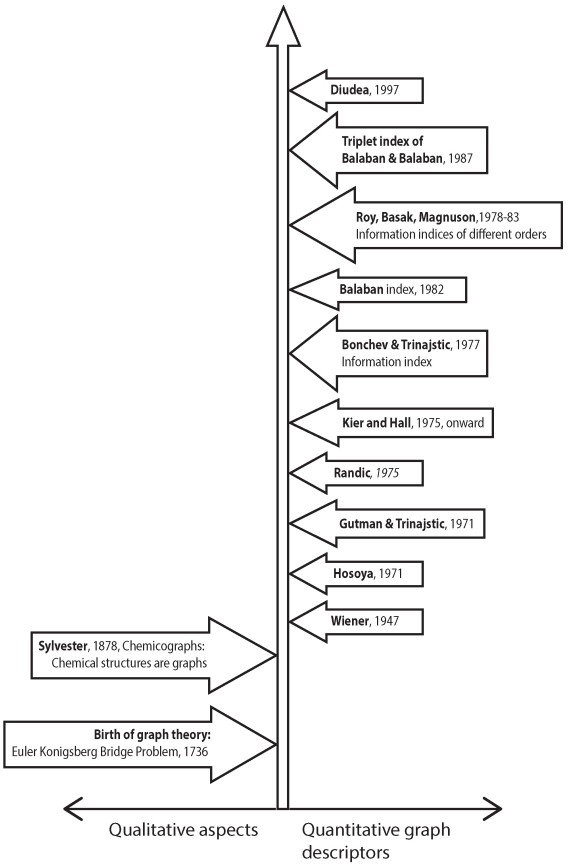

Figure 1. Graphical timeline of the development of the graph theory and its application in chemistry.

Trinajstić,[17] and Basak et al.[18] Figure 1 gives a bird's eye view of the development of graph theoretical molecular descriptors until now starting with the path-breaking discovery of graph theory by Euler.[19,20]

GRAPH THEORETICAL MATRICES AS THE SOURCE OF MOLECULAR DESCRIPTORS

Use of molecular descriptors and experimental properties in QSAR may be clearly understood through a formal exposition of the *structure-property similarity principle*—the central paradigm of SAR/QSAR.[21,22] Figure 2 represents an empirical property as a function α:C → R which maps the set C of compounds into the real line R. A non-empirical SAR may be looked upon as a composition of a description function β_1:C → D mapping each chemical structure of C into a space of non-empirical structural descriptors (D) and a prediction function β_2:D → R which maps the descriptors into the real line. When $[\alpha(C) - \beta_2\,\beta_1(C)]$ is within the range of experimental errors, we say that we have a good non-

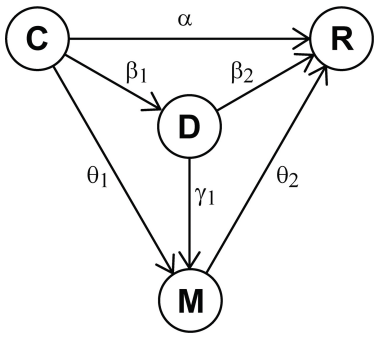

Figure 2. A schematic representation of methods used for QSAR formulation using the PAR and descriptor-based methods.

empirical predictive model. On the other hand, the property-activity relationship (PAR) is the composition of θ_1:C → M which maps the set C into the molecular property space M and θ_2:M → R mapping those molecular properties into the real line R. PAR seeks to predict one property (usually a complex property) of a molecule in terms of another (usually simpler) property or a set of properties. The latter group of properties may consist either of several experimentally determined quantities (e.g. melting point, boiling point, vapor pressure, partition coefficient) or substituent constants or solvato-chromic parameters (e.g. steric, electronic, hydrophobic, charge transfer substituent constants, hydrogen bond donor acidity, hydrogen bond acceptor basicity).

In the realm of chemical graph theory, description function β_1:C → D is often facilitated by various type of graph theoretical matrices, e.g., adjacency matrix, distance matrix, incidence matrix, D/D matrix, Laplacian matrix, *etc*.[23] In this process the molecular graphs are first converted to matrices and then numerical graph invariants are derived from the matrices using methods of matrix theory, information theory *etc*.

One of the less frequently used graph theoretical matrix is the Laplacian matrix,

$$\boldsymbol{L} = \boldsymbol{A} - \boldsymbol{D} \qquad (1)$$

where $\boldsymbol{A}$ is the adjacency matrix for the graph $G(V,E)$ with n vertices, while $\boldsymbol{D}$ is the degree matrix. $\boldsymbol{D}$ is a diagonal matrix with the degrees of the vertices in the main diagonal. The remaining elements of the matrix are zero.

The Laplacian matrix has been previously used in the field of mathematical chemistry for development of topological indices.[24–31] In this paper, we have developed a set of 54 new graph theoretical descriptors from the normalized Laplacian matrix. Subsequently, these newly developed descriptors were used for the formulation of QSAR models for the prediction of aquatic toxicity of a set of 69 benzene derivatives and solubility of a data set composed of diverse drug-like structures.

METHODS AND MATERIALS

The Database

The toxicity data, specifically the lethal concentration (96-hour LC_{50} in fathead minnow), for the 69 benzene derivatives were taken from Hall et al.[24] The data represent the acute aquatic toxicity (LC_{50}) measured in fathead minnow (*Pimephales promelas*). These data were compiled from eight other literature sources and include some original work conducted at the U.S. Environmental Protection Agency Environmental Research Laboratory (USEPA-ERL) in Duluth, MN, USA.

The solubility data set used in this study consists of 375 diverse drug-like structures. It was developed by combining a data set previously compiled by Rytting et al.[32] with a data set from Sirius Analytical Instruments,[33] referred to as the Rytting set and the Sirius set, respectively.

Computation of Convolution Based Molecular Descriptors

For the explanation of the descriptors presented in this article, it is important to note that the Laplacian matrix (***L***) is used to extract information from the properties of the vertices or edges of a chemical graph based on its topology. As the degrees of the vertices in a chemical graph may differ, it is preferable to use the symmetrically normalized Laplacian matrix ($\boldsymbol{L}_n$). The signless Laplacian matrix used here is defined as follows:

$$\boldsymbol{L} = \boldsymbol{D} + \boldsymbol{A} \tag{2}$$

The normalized Laplacian matrix is defined by:

$$\boldsymbol{L}_n = \boldsymbol{D}^{-1/2} \cdot \boldsymbol{L} \cdot \boldsymbol{D}^{-1/2} \tag{3}$$

In the previous equation, $\boldsymbol{D}^{-1/2}$ is defined as:

$$\boldsymbol{D}^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{d(1)}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{d(2)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{d(n)}} \end{bmatrix} \tag{4}$$

where $d(i)$ represents the vertex degree in the graph $G(V,E)$.

In order to proceed with the development of these descriptors, we need to define one more matrix called *property matrix* (***P***). The size of this matrix is $n \times k$. In this matrix, the number of rows (n) corresponds to the number of non-hydrogen atoms (or the number of vertices in the hydrogen-suppressed chemical graph), while the columns (k) correspond to properties of the isolated atoms, properties of atoms derived from their compounds or calculated properties derived from the chemical graph. The properties of the isolated atoms, which we use here, are: (1) atomic number, (2) relative atomic mass, (3) atomic radius, (4) first ionization potential, (5) electron affinity, (6) polarizability, (7) atomic volume and (8) van der Waals radius. The properties used for atoms derived from their compounds are: (9) electronegativity and (10) ionic radius. The only property derived from the chemical graph used in this work is (11) vertex degree. Therefore, for all molecules in the data set, the number of properties (columns) in the ***P*** matrix is 11.

In the next step, ***P*** is pre-multiplied by $\boldsymbol{L}_n$:

$$\boldsymbol{M} = \boldsymbol{L}_n \cdot \boldsymbol{P} \tag{5}$$

The resulting matrix (***M***) is called *Laplacian-transformed property matrix*. ***M*** has the same size as ***P***. Now, the values in columns of ***M*** do not represent the initial properties. The values in its columns are obtained as a result of "exchange of information" for the atomic property for the selected atom (vertex) and its neighbors at topological distance 1. Based on this, we can say that descriptors developed here are of distributive type because they encode information about the arrangement and interactions of vertices (atoms) and edges (bonds) within the molecular graph, not just their individual contributions.

The immediate neighborhood has the greatest influence on the selected atom. However, atoms at a topological distance of two or even three may also affect the behavior of the selected atom. If we wish to include the influence of atoms at a topological distance of two, we should pre-multiply ***P*** by $\boldsymbol{L}_n$ twice:

$$\boldsymbol{M} = \boldsymbol{L}_n \cdot \boldsymbol{L}_n \cdot \boldsymbol{P} \tag{6}$$

Similarly, the influence of the third neighborhood could be included in ***M*** if ***P*** is pre-multiplied three times by $\boldsymbol{L}_n$.

As previously stated ***M*** has the same size with ***P*** and it is shown here:

$$\boldsymbol{M} = \underbrace{\left.\begin{bmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1k} \\ m_{21} & m_{22} & m_{23} & \cdots & m_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{n1} & m_{n2} & m_{n3} & \cdots & m_{nk} \end{bmatrix}\right\}}_{\text{properties}} \text{atoms} \tag{7}$$

To extract descriptors from ***M***, a function processes the values in different columns. To test the predictive power of the descriptors obtained using this approach, we calculated them as either (1) mean values or (2) sums of the values in the selected column. Based on this, we can state

that the descriptors developed here are additive. A more detailed and systematic explanation of the procedure for calculating the proposed descriptors is provided in "Supporting Information 1".

For the eleven atomic properties mentioned earlier, using these two types of column processing, we obtained 22 descriptors (11 by averaging the columns and 11 by summing the columns of ***M***).

Additionally, based on this procedure, we also developed descriptors using line graphs (graphs in which the edges are treated as vertices). In this case, for the properties of the vertices, we used (1) differences between the properties of neighboring atoms and (2) sums of the properties of neighboring atoms. The reason for this is that, if we need to calculate descriptors representing the polarity of the molecule, the polarity of the chemical bonds should be considered. In that case, it makes sense to use the absolute value of the differences in electronegativity of the atoms forming the bond as a property of the vertices of the line graph. In other cases, for example, for a descriptor representing the polarizability of the molecule, it is better to define the properties for the vertices in the line graph as the sum of the polarizabilities of the neighboring atoms. Thus, when line graphs were used for the development of this type of descriptor, the properties of the bonds (columns) in matrix ***P*** were extracted using: (a) the sum of the values and (b) the absolute value of the differences of the properties of the individual atoms forming the bond. The properties used for calculation of these descriptors were: (1) relative atomic mass, (2) electronegativity, (3) electron affinity, (4) ionic radius, (5) atomic radius, (6) van der Waals radius, (7) polarizability, and (8) atomic volume. These two procedures for calculating the properties of the edges (sums and differences) in the line graphs, together with the properties mentioned in this section, yielded a total of 32 additional descriptors.

Using both approaches (atom-based and bond-based descriptors) presented here, we developed a total of 54 descriptors. These descriptors, along with the data sets used in this work, their abbreviations, and explanations of the abbreviations, are provided in "Supporting Information 2".

The entire procedure for the development of the convolution-based descriptors, as described here, is demonstrated in "Supporting Information 1" for a single molecule. The program is developed in Mathcad 15.[34]

QSAR Modeling for Exploration of Predictive Power of the Developed Descriptors

The predictive power of the developed descriptors was evaluated using multivariate linear regression (MLR) and counter-propagation artificial neural networks (CPANN). In-house software for MLR and CPANN was developed in the Matlab environment.[35] The software for CPANN is available on request.[36] The software for reading structures from SDF (structured-data format) was also developed in Matlab.[37] The software for calculating the new convolution-based descriptors,[38] discussed here, was also developed in the Matlab environment. For this study, the proposed descriptors were calculated using structures without hydrogen atoms.

Most of the properties of the atoms mentioned above have different physical units. To use only the numerical information stored in these properties after the convolution-based descriptors are calculated auto scaling or range scaling is required. In this paper, auto scaling was used before the QSAR modeling was performed.

RESULTS

The descriptors in this work were calculated using three pre-multiplications of ***P*** by the normalized Laplacian matrix. This approach was chosen to ensure that our descriptors included "interactions" among the vertices of the chemical graph at a maximum topological distance of three.

Multivariate Linear Regression Results

The search for the best models based on MLR started with univariate regression. After selecting the initial descriptor, MLR was used to select the remaining descriptors. The criteria for selecting descriptors were as follows:

- The first criterion was to select the descriptor that most improved the MLR model;
- The second criterion was that if, for example, electronegativity was chosen as an atomic property-based descriptor calculated as the mean of the column of ***M***, then the descriptor based on this electronegativity calculated as the sum of the column of ***M*** was excluded from further consideration. The same criterion was applied to the line graph-based descriptors.

Using this approach, we selected five descriptors for each data set, as shown in Table 1.

MLR Modeling of the Toxicity of the Benzene Derivatives

The data on the toxicity of the 69 benzene derivatives as well as the values for the previously mentioned descriptors are presented in Table 2.

The initial equation for the MLR model using auto scaled descriptors is:

$$\boldsymbol{y} = 0.699 \cdot \boldsymbol{b}_1 + 0.495 \cdot \boldsymbol{b}_2 + 1.321 \cdot \boldsymbol{b}_3 - 0.453 \cdot \boldsymbol{b}_4 - 0.367 \cdot \boldsymbol{b}_5$$

Here, ***y*** denotes the vector containing the modelled property, while $\boldsymbol{b}_i$ denotes the vector representing the *i*-th descriptor from Table 1. Potential collinearity among the

Table 1. The selected descriptors for the best MLR model. The first letter in the abbreviations stands for atomic (a_) or bond (b_) based descriptor.[a]

	#	Abbreviation	Description
Benzene derivatives data set	1	a_av_El_Affinity	Atom based descriptor – mean values of the Laplacian matrix based descriptor based on electron affinities
	2	a_av_Vertex_degree	Atom based descriptor – mean value of the Laplacian matrix based descriptor on vertex degrees
	3	a_su_Atom_vol	Atom based descriptor – sum of the values of the Laplacian matrix based descriptor based on atom volumes
	4	b_av_dif_vdW_radius	Bond based descriptor – Laplacian matrix based descriptor calculated using mean values of the differences of the van der Waals radii of the atoms which form bonds in the molecules
	5	b_av_sum_Polariz	Bond based descriptor – Laplacian matrix based descriptor calculated using mean values for the sums of the polarizabilities of the atoms which form bonds in the molecules
Solubility data set	#	Abbreviation	Description
	1	a_av_Ar	Atom based descriptor – mean values of the Laplacian matrix based descriptor derived from relative atomic weights
	2	a_av_El_Affinity	Atom based descriptor – mean values of the Laplacian matrix based descriptor derived from electron affinities
	3	a_av_Z	Atom based descriptor – mean values of the Laplacian matrix based descriptor derived from atomic number
	4	a_av_Atom_radius	Atom based descriptor – mean values of the Laplacian matrix based descriptor derived from atomic radii
	5	log P	Octanol-watter partion coefficient

[a] The second and third letter in the abbreviations represent whether the descriptor was calculated as mean value ('av') of the columns of the ***M*** or as a sum ('su'). For the bond-based descriptors the third part ('diff' or 'sum') represents whether difference or the sum of the atomic properties were used in order to calculate that descriptor. The remaining part of the abbreviations for all descriptors represent the atomic property on which they are based.

descriptors was assessed using the variance inflation factor (*VIF*). This parameter is calculated as the inverse of the correlation matrix for the descriptors.[39] The diagonal elements of the resulting inverse matrix correspond to the *VIF* values for each descriptor.[39] The *VIF* values obtained for this model are shown in Table 3. As a rule of thumb, variables (descriptors) with *VIF* > 10 should be further examined for collinearity. The only descriptor meeting this criterion is 'b_av_sum_Polariz'. In such cases, a *t*-test is used to check the statistical significance of this descriptor. At a significance level of α = 0.05, the absolute value calculated for the *t* statistic (*t* = –1.768) is not greater than the critical value ($t_{1-\alpha/2,v=63}$ = 1.998) which means that the null hypothesis (H_0), which states:

$$H_0 : a_5 = 0$$

should not be rejected. In the null hypothesis above, the parameter a_5 represents the regression coefficient (–0.367) corresponding to the 'b_av_sum_Polariz' descriptor. The *t*-test indicates that this descriptor should be excluded from the model. With this in mind, the final model for this data set is:

$$\boldsymbol{y} = 0.509 \cdot \boldsymbol{b}_1 + 0.444 \cdot \boldsymbol{b}_2 + 1.440 \cdot \boldsymbol{b}_3 - 0.453 \cdot \boldsymbol{b}_4$$

MLR Modeling of the Solubility Data Set

The solubility data set was also modeled using MLR. The resulting MLR model developed for the solubility data set is:

$$\boldsymbol{z} = -0.008 \cdot \boldsymbol{s}_1 + 0.059 \cdot \boldsymbol{s}_2 - 0.065 \cdot \boldsymbol{s}_3 - 0.050 \cdot \boldsymbol{s}_4 - 0.744 \cdot \boldsymbol{s}_5$$

Here, $\boldsymbol{z}$ is a vector representing the modeled property, and $\boldsymbol{s}_i$ are vectors representing the selected descriptors for this data set. The potential redundancy of the descriptors was also examined (Table 3). In this case, the calculated VIF values for all variables are below 10.

The expected versus predicted values for the modeled property using MLR with leave-one-out cross-validation, for both data sets are shown in Figure 3.

Validation of the Best MLR Models for Both Data Sets

The root-mean-square error of prediction (*RMSEP*) for the developed model for the benzene derivatives data set is 0.376 (Figure 3a) while the correlation coefficient between expected and found values for the modeled property is 0.871. The predictive power of this model was evaluated using leave-one-out cross-validation (Figure 3b), which is a widely accepted validation procedure in QSAR modeling (e.g. Refs. [40,41]). Correlation coefficient for cross-validation is 0.840, while root-mean-square error of cross-validation is 0.412. At the same time, q^2 for this model is 0.670. Similar results, but with more parameters, were previously obtained on this data set of chemicals earlier using MLR models with three topological and two physicochemical descriptors (the root-mean-square error of cross-validation was 0.312[40]).

Table 2. The values for the selected descriptors used for modeling of the toxicity of the benzene derivatives.

		1	2	3	4	5	
		a_av_El_Affinity	a_av_Vertex_degree	a_su_Atom_vol	b_av_dif_vdV_radius	b_av_sum_Polariz	$-\log(LC_{50})$
1	benzene	10.0800	80.0000	219.8400	0.0000	28.8000	3.4
2	bromobenzene	11.6218	74.5776	372.9090	7.6539	29.9313	3.89
3	chlorobenzene	11.8277	74.5776	322.7372	2.5513	29.0116	3.77
4	phenol	10.0565	78.8749	306.0133	9.1847	27.5736	3.51
5	toluene	9.8917	78.8749	251.6894	0.0000	28.6028	3.32
6	1,2-dichlorobenzene	13.0852	78.0689	423.4035	4.2861	29.0651	4.4
7	1,3-dichlorobenzene	13.1298	78.0005	425.2169	4.4266	29.2467	4.3
8	1,4-dichlorobenzene	13.1412	78.0053	425.7116	4.4648	29.2227	4.62
9	2-chlorophenol	11.5596	78.0689	406.9406	9.8581	27.8609	4.02
10	3-chlorotoluene	11.4421	78.0005	354.4356	2.2133	28.8943	3.84
11	4-chlorotoluene	11.4472	78.0053	354.6638	2.2324	28.8666	4.33
12	1,3-dihydroxybenzene	10.0417	78.0005	391.8945	15.9356	26.7679	3.04
13	3-hydroxyanisole	10.1318	72.2985	454.2196	21.8544	25.8033	3.21
14	2-methylphenol	9.8920	78.0689	337.0016	7.7151	27.5186	3.77
15	3-methylphenol	9.8981	78.0005	337.7744	7.9678	27.6550	3.29
16	4-methylphenol	9.8973	78.0053	337.9399	8.0366	27.6138	3.58
17	4-nitrophenol	8.8622	73.3950	645.8832	15.7330	24.5827	3.36
18	1,4-dimethoxybenzene	10.2002	65.0865	516.6786	26.6793	24.9773	3.07
19	1,2-dimethylbenzene	9.7501	78.0689	283.5256	0.0000	28.3804	3.48
20	1,4-dimethylbenzene	9.7532	78.0053	283.6160	0.0000	28.5104	4.21
21	2-nitrotoluene	8.7565	69.0260	590.7731	9.1569	25.2651	3.57
22	3-nitrotoluene	8.7517	68.9771	591.5224	9.2713	25.3458	3.63
23	4-nitrotoluene	8.7468	73.3950	591.5593	9.3037	25.3000	3.76
24	1,2-dinitrobenzene	8.1036	64.9791	897.5127	15.1968	23.2217	5.45
25	1,3-dinitrobenzene	8.0831	64.9700	899.3907	15.4522	23.2192	4.38
26	1,4-dinitrobenzene	8.0759	68.7829	899.5027	15.5062	23.1598	5.22
27	2-methyl-3-nitroaniline	8.1479	69.3159	694.6452	12.9684	25.0187	3.48
28	2-methyl-4-nitroaniline	8.1398	66.6608	695.6325	13.1042	25.0782	3.24
29	2-methyl-5-nitroaniline	8.1421	63.0381	695.3574	13.0625	25.0844	3.35
30	2-methyl-6-nitroaniline	8.1611	69.3159	693.0864	12.7547	25.0392	3.8
31	3-methyl-6-nitroaniline	8.1547	73.2712	694.1394	12.9086	25.1355	3.8
32	4-methyl-2-nitroaniline	8.1559	69.2568	694.3394	12.9208	25.1496	3.79
33	4-hydroxy-3-nitroaniline	8.2457	69.2568	748.8402	18.6750	24.5164	3.65
34	4-methyl-3-nitroaniline	8.1434	69.2568	695.8065	13.1239	25.1302	3.77
35	1,2,3-trichlorobenzene	14.0575	77.4290	523.7742	5.6253	29.1408	4.89
36	1,2,4-trichlorobenzene	14.1049	77.3621	525.9474	5.7603	29.2995	5
37	1,3,5-trichlorobenzene	14.1348	77.2933	527.2790	5.8511	29.4918	4.74
38	2,4-dichlorophenol	12.7487	77.3621	509.4845	10.7131	28.2327	4.3
39	3,4-dichlorotoluene	12.6047	77.3621	455.1660	3.7929	28.9873	4.74
40	2,4-dichlorotoluene	12.6226	77.3621	456.0085	3.8553	28.9962	4.54
41	4-chloro-3-methylphenol	11.2557	77.3621	439.6137	8.9874	27.9015	4.27
42	2,4-dimethylphenol	9.7719	77.3621	369.0306	6.8578	27.6208	3.86
43	2,6-dimethylphenol	9.7659	77.4290	368.0496	6.5902	27.5088	3.75
44	3,4-dimethylphenol	9.7734	77.3621	369.6747	7.0825	27.5983	3.9
45	2,4-dinitrophenol	8.2324	64.2533	983.6289	18.8476	23.0627	4.04
46	1,2,4-trimethylbenzene	9.6458	77.3621	315.5546	0.0000	28.3843	4.21
47	2,3-dinitrotoluene	8.1665	63.6804	928.6627	13.9111	23.5834	5.01
48	2,4-dinitrotoluene	8.1457	64.2533	930.5953	14.1506	23.5821	3.75
49	2,5-dinitrotoluene	8.1428	67.1570	930.6704	14.1755	23.5619	5.15
50	2,6-dinitrotoluene	8.1503	67.0562	929.8164	14.0652	23.5179	3.99
51	3,4-dinitrotoluene	8.1610	67.1115	929.4666	14.0028	23.6467	5.08
52	3,5-dinitrotoluene	8.1461	64.2289	931.3205	14.2136	23.6894	3.91
53	1,3,5-trinitrobenzene	7.6945	61.6686	1239.1137	18.4777	22.2313	5.29
54	2-methyl-3,5-dinitroaniline	7.7177	64.8248	1034.2196	16.7327	23.5707	4.12
55	2-methyl-3,6-dinitroaniline	7.7263	67.5452	1032.0520	16.5372	23.4797	5.34
56	3-methyl-2,4-dinitroaniline	7.7274	62.7835	1032.5011	16.5825	23.5104	4.26
57	5-methyl-2,4-dinitroaniline	7.7245	64.8140	1033.2267	16.6368	23.5904	4.92
58	4-methyl-2,6-dinitroaniline	7.7381	64.8051	1031.6678	16.4662	23.6275	4.21
59	5-methyl-2,6-dinitroaniline	7.7390	64.8255	1030.7589	16.3999	23.5286	4.18
60	4-methyl-3,5-dinitroaniline	7.7185	62.6017	1034.6020	16.7756	23.5992	4.46
61	2,4,6-tribromophenol	13.2443	76.8360	757.6315	21.3512	30.2926	4.7
62	1,2,3,4-tetrachlorobenzene	14.8369	76.9022	624.2005	6.6967	29.2231	5.43
63	1,2,4,5-tetrachlorobenzene	14.8772	76.8339	626.2346	6.7967	29.3776	5.85
64	2,4,6-trichlorophenol	13.6695	76.8360	609.6858	11.0712	28.4654	4.33
65	2-methyl-4,6-dinitrophenol	8.2763	59.6182	1014.6542	17.2958	23.4401	5
66	2,3,6-trinitrotoluene	7.7856	64.6439	1267.7617	17.0077	22.4994	6.37
67	2,4,6-trinitrotoluene	7.7743	57.8402	1269.5908	17.1617	22.5472	4.88
68	2,3,4,5-tetrachlorophenol	14.3651	76.4553	707.8253	11.5455	28.4676	5.72
69	2,3,4,5,6-pentachlorophenol	14.9646	76.1321	806.0487	11.7095	28.5241	6.06

Table 3. The values for the variance inflation factor for the two data sets.

Benzene derivatives data set					
#	1	2	3	4	5
Abbreviation	a_av_El_Affinity	a_av_Vertex_degree	a_su_Atom_vol	b_av_dif_vdW_radius	b_av_sum_Polariz
VIF	8.2593	8.5340	7.5035	2.7935	21.5638
Solubility data set					
#	1	2	3	4	5
Abbreviation	a_av_Ar	a_av_El_Affinity	a_av_Z	a_av_Atom_radius	logP
VIF	4.082	1.1762	4.052	1.480	1.015

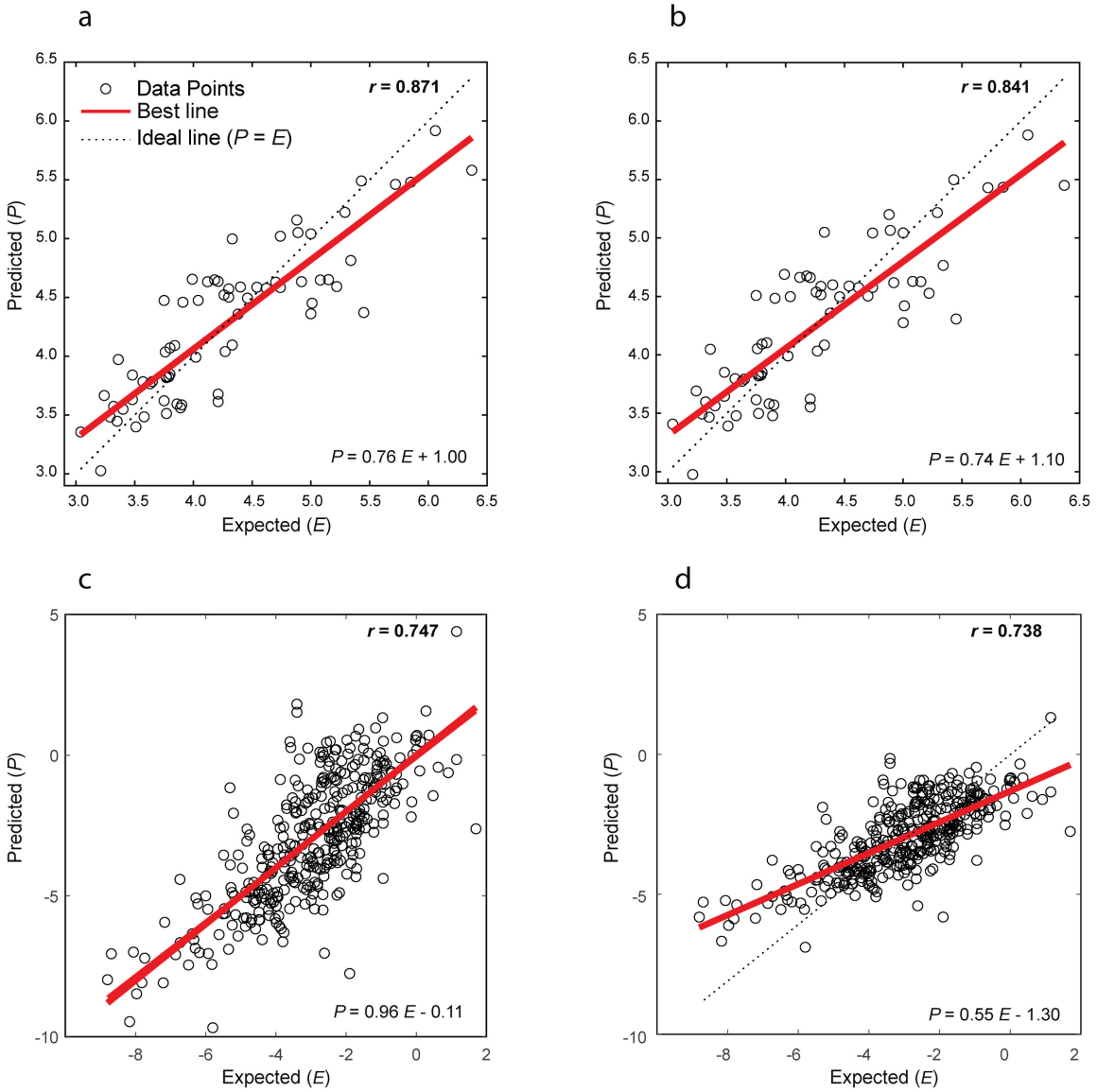

Figure 3. Expected vs. predicted values for the MLR models obtained using benzene derivative data set (a – training set; b – leave-one-out cross-validation) and for the solubility data set (c – training set; d – 20-fold cross-validation).

In addition, *y*-randomization was used to assess whether the model performances were due to chance correlation. To this end, the entire procedure was repeated 10,000 times. The average *RMSEP* for cross-validation obtained from *y*-randomization is 0.792, with a standard deviation of 0.020. Comparison of the *RMSEP* for cross-validation (0.412) with that from *y*-randomization (0.789) shows that the cross-validation results are approximately 19 standard deviations from the mean of the *y*-randomization distribution. This indicates that the MLR model is not the result of chance correlation.

The best MLR model for predicting solubility had an *RMSEP* of 0.263, while the *RMSEP* from 20-fold cross-validation was 0.402. The q^2 value obtained by the cross-validation was 0.628. In this case, *y*-randomization was also applied under the previously conditions described. The *RMSEP* obtained for this procedure is 1.01 with standard deviation of 0.005. This means that the model presented here for prediction of the solubility is not the result of a chance correlation.

Based on the visual inspection of Figure 3 and the previously discussed quantitative parameters, we conclude that the MLR models developed for both data sets using these new Laplacian matrix-based descriptors show good predictive performance.

Counter-propagation Artificial Neural Networks Results

In a search for better results compared to those obtained using MLR, we tried to develop nonlinear models based on counter-propagation artificial neural networks.[12,13] Unlike most of the artificial neural networks, which are considered as "black box" models, CPANN could help in understanding the influence of the descriptors on the model. In some cases, where the interpretation of the descriptors is straightforward, it could help in development of mechanistically interpretable models.[42,43]

Here, the descriptors previously selected for MLR modeling (Table 1) were also used to optimize the models based on CPANN. To perform CPANN modeling automatically and to develop simpler models, we used a genetic algorithm (1) to select the best descriptors for the models and (2) to automatically search for the optimal network size and number of training epochs.[44] In addition, to further simplify the model, we used our algorithm for automatic adjustment of the relative importance of the descriptors.[44] This algorithm usually produces simpler models that are easier to interpret.[44]

Due to the small number of selected descriptors, the genetic algorithm converged relatively quickly to good CPANN models for both data sets.

Modeling of the Toxicity of the Benzene Derivatives Using CPANN

Only two descriptors were selected by the genetic algorithm for modeling the $-\log(LC_{50})$ of the benzene derivatives data set. Both descriptors are based on the properties of the isolated atoms. The descriptor with a relative importance of 1.000 is based on the convolution of the electron affinities of the atoms (a_av_El_Affinity). The second descriptor, with a lower relative importance (0.042), is based on the sums of the atomic volumes (a_su_Atom_vol) of the elements. The weight levels for CPANN in this model are shown in Figure 4.

A simple comparison of the weight levels (in Figure 4) shows that the selected descriptors (a_av_El_Affinity and a_su_Atom_vol) create two different regions with high values for the modeled property. The descriptor based on electron affinity helps in clustering the structures with high values for $-\log(LC_{50})$ in the upper left corner of the CPANN. The 'a_su_Atom_vol' descriptor groups the remaining structures with high values of the

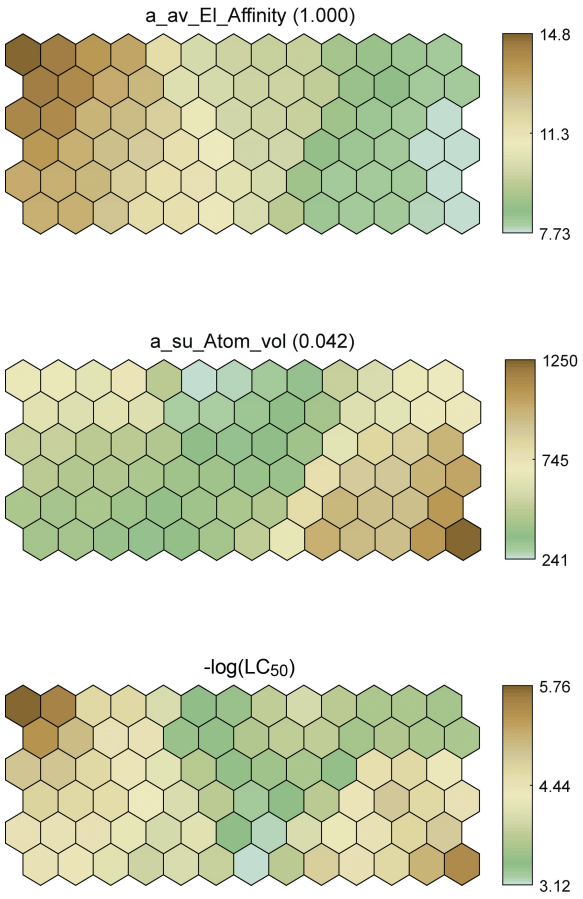

Figure 4. The weights levels for one of the best CPANN models based on the convolutional descriptors presented here.

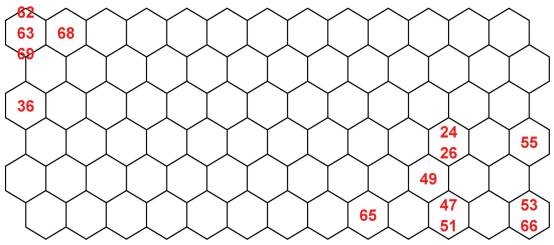

Figure 5. Trained and labeled CPANN (for the model show in Figure 4) with numbers of the structures with the values for the modeled property larger than 80th percentile.

modeled property in the lower right corner of the map. It is interesting to see whether there is a difference between the structures mapped in these two regions of the CPANN. For this purpose, the structures with values for the modeled property larger than the 80th percentile were mapped on the trained CPANN. The CPANN labeled with these structures is shown in Figure 5.

The structures in the upper left corner of the CPANN (Figure 5) are those that have three (structure number 36), four (structures: 62, 63 and 68) and five (number 69) chlorine atoms attached to the benzene ring. In the lower right corner of the CPANN, nitro substituted benzene derivatives are mapped. Among these structures two $–NO_2$ groups are present in the molecules labeled with 24, 25, 47, 49, 51, 55 and 64 on Table 1 together with structures 53 and 55 that have three nitro groups (see "Supporting Information 3"). The experimental versus predicted values for $–\log(LC_{50})$ for this model are presented in Figure 6a and 6b.

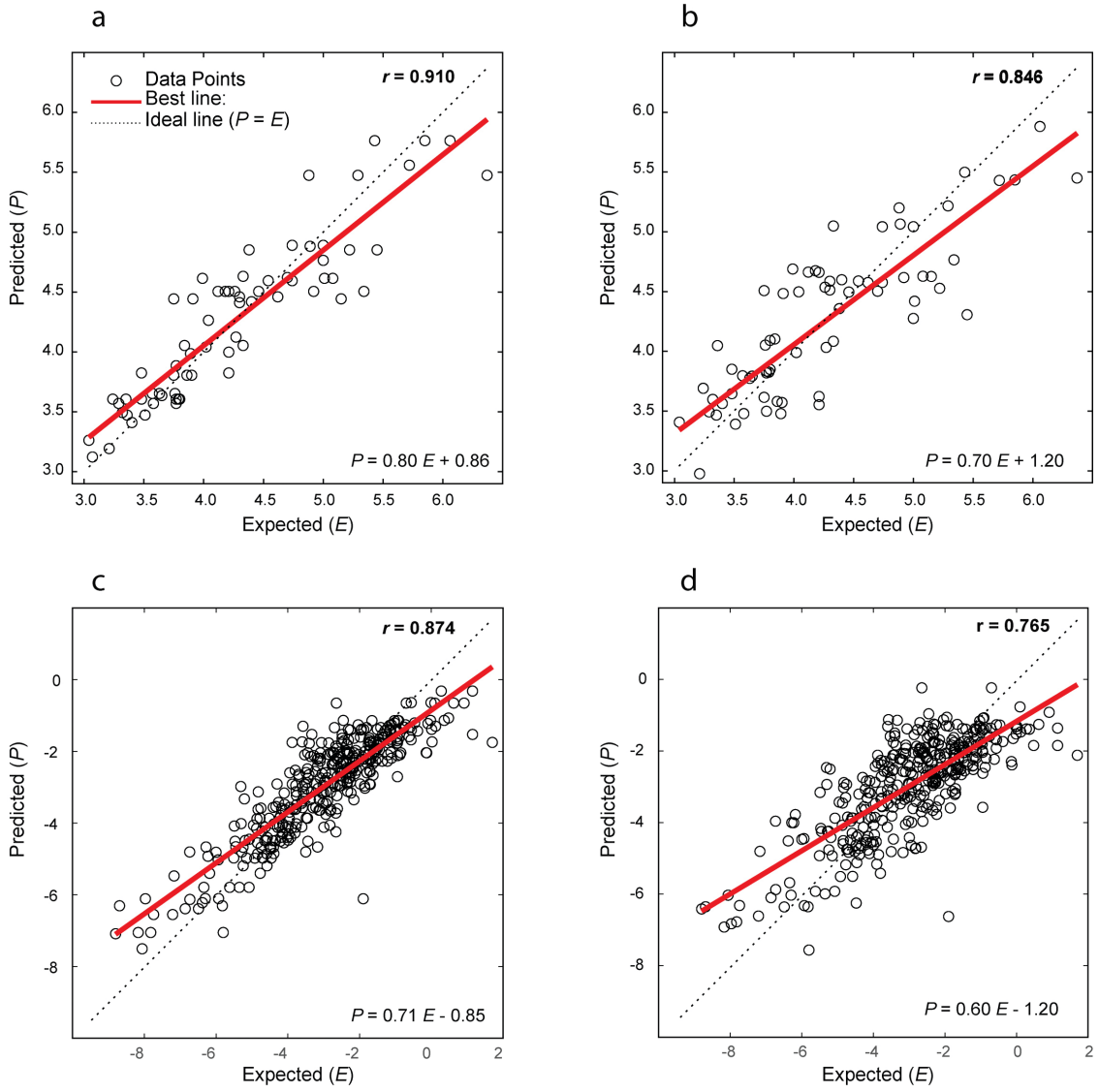

Figure 6. Expected *vs*. found values for the modeled property using CPANN (a) for the training set and (b) for cross-validation for the benzene derivatives data set and (c) for the training set and (d) cross-validation for the solubility data set.

The root mean square error of prediction for the training set of this model is 0.414. The correlation coefficient for the training set is 0.910, while the correlation coefficient for leave-one-out cross-validation is 0.846. The q^2 for cross-validation is 0.713, and the root mean square error for cross-validation is 0.532. The numerical parameters for the presented CPANN model indicate good generalization performances. Performance was examined using *y*-randomization. In this case, the mean value of the performance (after 10,000 repetitions) with cross-validation was 1.131, with a standard deviation of 0.061. Based on this, we conclude that the performance of our models (based on leave-one-out cross-validation) cannot be attributed to chance correlation. Here, the *RMSEP* for cross-validation is 4.7 standard deviations lower than the average performance of the random models.

Solubility Modeling by CPANN

For modelling solubility, the octanol-water partition coefficient (log P) plays a very important role. Therefore, log P was included among the selected descriptors (Table 1). The remaining descriptors were chosen using a stepwise addition procedure. In this case, all descriptors are derived from atomic properties. Three are size-based descriptors (a_av_Ar, a_av_Z, and a_av_Atom_radius). The fourth is an electronic descriptor derived from electron affinity (a_av_El_Affinity).

After several repetitions of the optimization using a genetic algorithm, multiple models with the same subset of descriptors and comparable performance were identified. These models were based on only three descriptors. The selected descriptors are a_av_At, a_av_El_Affinity, and logP (see Figures 6c and 6d). The relative importance of these descriptors during the search for the winning neuron was determined using our procedure for automatic adjustment of relative importance.[44] The relative importance of these descriptors was:

$$a\_av\_At : a\_av\_El\_Affinity : \log P = 1.000 : 0.042 : 0.750$$

Unlike when MLR was used here, probably because the algorithm can model nonlinearities, the log P descriptor does not have a greater influence on the mapping of the structures on the CPANN.

The root mean square error of prediction for the training set is 0.489, while for the 20-fold cross-validation it is 0.644. This model has a q^2 value for the cross-validation of 0.585. In addition, *y*-randomisation was performed for the model under the same conditions as previously described. Here, the arithmetic mean obtained for the *RMSEP* is 1.08 with a standard deviation of 0.066. Considering the standard deviation for this distribution and the *RMSEP* obtained for the CPANN model after cross-validation, we can conclude that our results are not due to chance correlation.

As presented, the results obtained using CPANN and MLR are comparable. Based on the cross-validation results, MLR appears to show better generalization performance in this case. Probably due to the non-linear nature of the CPANN algorithm, only three of the five preselected descriptors are included in the final model.

DISCUSSION

The results presented in this article indicate that Laplacian-based convolution descriptors could become a valuable tool in future SAR/QSAR studies. In this case, only a few atomic properties were used to demonstrate the predictive power of this approach.

The list of developed descriptors can be extended by incorporating additional properties of individual atoms or bonds. For example, additional descriptors can be developed based on the line graph, using bond orders. For atom-based descriptors, it is also possible to develop descriptors based on the hybridization states of the atoms in the molecule.

The descriptors developed here are of distributive and additive types. The Laplacian matrix enables information exchange between neighboring atoms or, in the case of line graphs, between neighboring chemical bonds. Therefore, this new group of descriptors is distributive. Additionally, extraction of the descriptors from the columns of the Laplacian-transformed property matrix (***M***) involves either (1) summing the values in the columns or (2) calculating their mean values. Based on this, we can state that these descriptors are additive.

One very important detail which must be considered is that sometimes it is not desirable to use the descriptors obtained using only one pre-multiplication of the property matrix (***P***) with the normalized Laplacian ($\boldsymbol{L}_n$). We noticed that, for these data sets, if the descriptors were calculated using only one pre-multiplication by $\boldsymbol{L}_n$ most of the atom-based descriptors will have same values for meta and para substituted benzene derivatives. This means that at least two pre-multiplications with $\boldsymbol{L}_n$ are required to characterize the inductive effect of atoms at a topological distance of two in order to have an effect, and, in this case, to help distinguish between meta- and para-substituted benzene derivatives. Another reason to interpret this as an influence of the inductive effect is that it was not observed with ortho-substituted benzene derivatives. Here, only one per-multiplication with the Laplacian matrix separates ortho-substituted benzene derivatives from meta- and para-substituted benzenes.

The benzene data set, although apparently congeneric, exhibits considerable structural diversity, which has mechanistic implications for toxicity.

In Table 2, compounds 18, 19, and 20 (1,4-dimethoxybenzene, 1,2-dimethylbenzene, and 1,4-dimethylbenzene) are highly hydrophobic molecules with low dipole moment values. For example, the dipole moment of 1,4-dimethoxybenzene is approximately 1.70 Debye in solution.

In contrast, compounds 47, 48, and 49 (2,3-dinitrotoluene, 2,4-dinitrotoluene, and 2,5-dinitrotoluene) are considerably more polar. The dipole moment of 2,3-dinitrotoluene is approximately 3.45 D, while 2,4-dinitrotoluene has a dipole moment of about 4.35 D. These chemicals act as non-polar narcotics. Dinitrotoluenes are nitroaromatic compounds that are sparingly soluble in water, a key characteristic of non-polar substances. The molecules consist largely of a non-polar benzene ring and a methyl group, and although the nitro groups add some polarity to specific bonds, the overall molecular structure is predominantly non-polar. The similar compound 1,4-dinitrobenzene is non-polar because its symmetry cancels out the dipole moments. The theoretical dipole moment of 1,4-dinitrobenzene (compound No. 26) is zero because the two nitro group dipoles are oriented in opposite directions and cancel each other out due to the molecule's symmetry.

Let us now consider the pK_a values of the phenols of Table 2:

64	2,4,6-trichlorophenol	≈ 6.0 – 6.23
65	2-methyl-4,6-dinitrophenol	≈ 4.42
68	2,3,4,5-tetrachlorophenol	≈ 6.15 – 6.35
69	2,3,4,5,6-pentachloropheno	≈ 4.7 – 5.3

In aqueous solution at 25 °C, the pK_a of the mitochondrial oxidative phosphorylation uncoupler 2,4-dinitrophenol (2,4-DNP) is approximately 4.1. Based on this, it can be assumed that the four phenolic compounds mentioned above will act as uncouplers in the fathead minnow. Therefore, it is reasonable that a genetic algorithm-based QSAR model selected one electron affinity-based Laplacian descriptor and one size-dependent index.

For aqueous solubility, calculated log P and four Laplacian descriptors were required to develop a good predictive model. The electron affinity-based Laplacian descriptor is electronic in nature (see Table 1). The other three Laplacian indices encode information about molecular size and shape. The critical role of hydrophobicity in predicting solubility is well established in the QSAR literature.[45]

CONCLUSION

In this paper, we report the development of a set of novel Laplacian matrix-based molecular descriptors. A limited selection of these descriptors has been applied to QSAR studies of two data sets: (1) aquatic toxicity of benzene derivatives and (2) aqueous solubility of another collection of molecules. The results indicate that the newly formulated Laplacian descriptors encode sufficiently valuable structural information to be useful as new QSAR descriptors. Following the 'diversity begets diversity' principle,[46] it would be desirable to develop and explore the utility of Laplacian class of chemodescriptors in the classification and property (as well as bioactivity prediction) of larger and structurally diverse databases. Such studies are currently in progress, the results of which will be reported in the future communications.

Supplementary Information. Supporting information to the paper is attached to the electronic version of the article at: https://doi.org/10.5562/cca4199.

PDF files with attached documents are best viewed with Adobe Acrobat Reader which is free and can be downloaded from Adobe's web site.

REFERENCES

[1] S. C. Basak, *Drug Dev. Res*. **2011**, *72*, 225–233. https://doi.org/10.1002/ddr.20428

[2] D. M. Hawkins, S. C. Basak, X. Shi, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 663–670. https://doi.org/10.1021/ci0001177

[3] D. M. Hawkins, S. C. Basak, D. Mills, *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 579–586. https://doi.org/10.1021/ci025626i

[4] X. Guo, M. Randić, S. C. Basak, *Chem. Phys. Lett*. **2001**, *350*, 106–112. https://doi.org/10.1016/S0009-2614(01)01246-5

[5] M. Randić, X. Guo, S. C. Basak, *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 619–626. https://doi.org/10.1021/ci000120q

[6] M. Vračko, S. C. Basak, K. Geiss, F. Witzmann, *J. Chem. Inf. Model*. **2006**, *46*, 130–136. https://doi.org/10.1021/ci0502597

[7] L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, CA, **1999**.

[8] D. Bonchev; N. Trinajstić, *J. Chem. Phys*. **1977**, *67*, 4517–4533. https://doi.org/10.1063/1.434593

[9] S. C. Basak, M. Vračko, (Eds), *Big Data Analytics in Chemoinformatics and Bioinformatics: With Applications to Computer-Aided Drug Design, Cancer Biology, Emerging Pathogens and Computational Toxicology*, Elsevier, 1st Edition, **2022**. https://doi.org/10.1016/C2020-0-02815-X

[10] S. C. Basak (Ed.), *Mathematical Descriptors of Molecules and Biomolecules* (*Applications in Chemistry, Drug Design, Chemical Toxicology, and Computational Biology*), Synthesis Lectures on Mathematics & Statistics (SLMS), Springer Nature, **2024**. https://doi.org/10.1007/978-3-031-67841-7

[11] S. Majumdar, S. C. Basak, C. N. Lungu, M. V. Diudea, G.D. Grunwald, *Mol. Inf.* **2019**, *38*, 1800164. https://doi.org/10.1002/minf.201800164
[12] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*. WCH: Weinheim, **1999**.
[13] M. Novič, J. Zupan, *J. Chem. Inf. Comput .Sci*. **1995**, *35*, 454–466. https://doi.org/10.1021/ci00025a013
[14] H. Wiener, *J. Am. Chem. Soc*. **1947**, *69*, 17–20. https://doi.org/10.1021/ja01193a005
[15] A. T. Balaban, *Chem. Phys. Lett*. **1982**, *89*, 399–404. https://doi.org/10.1016/0009-2614(82)80009-2
[16] M. Randić, *J. Am. Chem. Soc*. **1975**, *7*, 6609–6615. https://doi.org/10.1021/ja00856a001
[17] N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, Florida, USA, **1992**.
[18] S. C. Basak, G. Restrepo, J. L. Villaveces (Eds). *Advances in Mathematical Chemistry and Applications*, Vol. 1 & 2, Elsevier-Bentham, **2015**. https://doi.org/10.2174/9781681080529115020l
[19] S. C. Basak, *Croat. Chem. Acta,* **2020**, *93*, 247–258. https://doi.org/10.5562/cca3759
[20] I. Euler, *Comment. Acad. Sci. U. Petrop*. **1736**, *8*, 128–140.
[21] M. Johnson, S. C. Basak, G. A. Maggiora, *Math. Comp. Model.*, **1988**, *11*, 630–634. https://doi.org/10.1016/0895-7177(88)90569-9
[22] S. C. Basak, G. J. Niemi, G. D. Veith, *J. Math. Chem.*, **1991**, *7*, 243–272. https://doi.org/10.1007/BF01200826
[23] D. Janežič, A. Miličević, S. Nikolić, N. Trinajstić, *Graph-Theoretical Matrices in Chemistry*, CRC Press, Boca Raton, Florida, USA, **2015**. https://doi.org/10.1201/b18389
[24] L. H. Hall, L. B. Kier, G. Phipps, *Environ. Toxicol. Chem.*, **1984**, *3*, 355–365. https://doi.org/10.1002/etc.5620030301
[25] W. C. Herndon, M. L. Ellzey, Jr., *MATCH Commun. Math. Comput. Chem*. **1986**, *20*, 53–79.
[26] B. Mohar, *Graphs Comb*. **1997**, *7*, 53–64. https://doi.org/10.1007/BF01789463
[27] B. Mohar, D. Babić, N. Trinajstić, *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 153–154. https://doi.org/10.1021/ci00011a023
[28] N. Trinajstić, D. Babić, S. Nikolić, D. Plavšić, D. Amić, Z. Mihalić, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 368–376.
[29] S. Marković, I. Gutman, Ž. Bančević, *J. Serb. Chem. Soc*. **1995**, *60*, 633–636.
[30] R. B. Mallion, N. Trinajstić, *MATCH Commun. Math. Comput. Chem*. **2003**, *48*, 97–116.
[31] C. Kirby, D. J. Klein, R. B. Mallion, P. Pollak, H. Sachs, *Croat. Chem. Acta*, **2004**, *77*, 263–278.
[32] E. Rytting, K. A. Lentz, X. Q. Chen, F. Qian, S. Venkatesh, *The AAPS J*. **2005**, *7*, E78–E105. https://doi.org/10.1208/aapsj070110
[33] K. J. Box, J. E. A. Comer. *Curr. Drug Metab.* **2008**, *9*, 869–878. https://doi.org/10.2174/138920008786485155
[34] Mathcad 15.0, 1986–2003, PTC.
[35] MATLAB 6.5, 1984–1998 Mathworks.
[36] I. Kuzmanovski, M. Novič, *Chemom. Intell. Lab. Syst*. **2008**, *90*, 84–91. https://doi.org/10.1016/j.chemolab.2007.07.003
[37] I. Kuzmanovski, *SDF reader for Matlab*, Version 1.0, unpublished software.
[38] I. Kuzmanovski, *MolConvol*, Version 1.0, unpublished software.
[39] D. A. Belsley, E. Kuh, R. E. Welsh, *Regression Diagnostics*. New York, NY: John Wiley & Sons, Inc., **1980**. https://doi.org/10.1002/0471725153
[40] S. C. Basak, B. D. Gute, B. Lučić, S. Nikolić, N. Trinajstić, Computers & Chemistry 2000, 24, 181–191. https://doi.org/10.1016/S0097-8485(99)00059-5
[41] A. Euldji, M. Laidi, M. Hentabli, A. Madani, S. Hanini, *Croat. Chem. Acta*, **2025**, *98*, 15–26. https://doi.org/10.5562/cca4137
[42] N. Stojić, S. Erić, I. Kuzmanovski, *J. Mol. Graphics Modell*. **2010**, *29*, 450–460. https://doi.org/10.1016/j.jmgm.2010.09.001
[43] G. Stojković, M. Novič, I. Kuzmanovski, *Chemom. Intell. Lab. Syst*. **2010**, *102*, 123-129. https://doi.org/10.1016/j.chemolab.2010.04.013
[44] I. Kuzmanovski, M. Novič, M. Trpkovska, *Anal. Chim. Acta*, **2009**, *642* 142–147. https://doi.org/10.1016/j.aca.2009.01.041
[45] C. Hansch, J. E. Quinlan, G. L. Lawrence, *J. Org. Chem.* **1968**, *33*, 347–350. https://doi.org/10.1021/jo01265a071
[46] S. C. Basak, S. Majumdar, *Comput. Drug Des*. **2016**, *12*, 84–86. https://doi.org/10.2174/1573409912021607131904l6