Revue Roumaine de Chimie, **1999**, *44* (6), 619–631

ARTIFICIAL NEURAL NETWORKS APPLICATIONS. PART 11.[1]

MolNet PREDICTION OF ALKANE DENSITIES

Ovidiu IVANCIUC

"Politehnica" University, Department of Organic Chemistry,
Faculty of Industrial Chemistry, Oficiul 12 CP 243,
78100 Bucharest, Roumania
E-mail: o_ivanciuc@chim.upb.ro

Received March 6, 1998

MolNet, a new multy-layer feedforward neural network is presented together with its application to the computation of alkane densities. The MolNet neural network changes its topology (the number of neurons in the input and hidden layers, together with the number and type of connections) according to the molecular structure of the chemical compound presented to the network. The structure of each molecule is encoded in the corresponding molecular graph that is used to set the MolNet topology. Three structural descriptors derived from the molecular graph are used as input data for the first layer of neurons, namely the degree, the distance sum, and the reciprocal distance sum.

INTRODUCTION

The high interest in the application of Artificial Neural Networks (ANN)[2,3] in chemistry,[4] chemical engineering,[5] and biochemistry,[6] is a consequence of their high flexibility in modeling nonlinear relationships. Various physico-chemical properties of inorganic and organic compounds were predicted in Quantitative Structure-Property Relationships (QSPR) studies involving neural networks.

An important problem for the QSPR applications of neural models is the numerical representation of the chemical structure by various constitutional, topological, geometric, or quantum descriptors. A large number of structural descriptors were used as input to neural networks in QSPR studies applying Multi-Layer Feedforward (MLF) neural models: connection table describing the substituents;[7] modified bond-electron matrix containing as structural information the formal bond order between a pair of atoms and the atomic number Z;[8] number of bonds between methyl groups;[9] distance degree sequence in molecular graphs;[10] constitutional descriptors and topological indices;[11] numerical code;[12] counts of various molecular subgraphs (clusters);[13] vectorial representation of the chemical structure of the substituents;[14] topo-stereochemical code describing the environment of an atom;[15,16] the three-dimensional structure encoded in the 3D MORSE (Molecule Representation of Structures based on Electron Diffraction) representation;[17,18] atom type electrotopological state;[19] presence of a substituent (coded with 1) or absence (coded with 0);[20] topological autocorrelation vectors;[21] molecular similarity matrices.[22,23]

In structure-property studies an MLF network receives information related to the molecular structure of the chemical compounds only from the input neurons, and the topology of the neural network (the number of neurons in the input, hidden, and output layers, together with connections between them) is constant for all molecules presented to the network. Three new neural models, that encode into their topology the molecular structure of each compound, were recently introduced: the ChemNet proposed by Kireev;[24] the Baskin, Palyulin, Zefirov (BPZ) neural device;[25] and the MolNet defined by

Ivanciuc.[1,26,27] The above three neural models use a set of rules to build the network according to the chemical structure of each molecule examined by the ANN. MolNet uses various atomic descriptors as input structural parameters, and it was successfully applied for the computation of a large number of molecular properties of organic compounds. In the present investigation we use MolNet to compute alkane densities.

MolNet Description

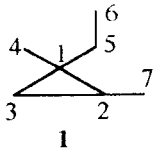
MolNet is an MLF neural network designed to compute molecular properties using information on the chemical structure. The MolNet topology, *i.e.* the number of units in each layer and the connections between layers, changes according to the structure of each molecule presented to the network. Each atom from the molecular graph of a chemical compound has a corresponding unit in the input and hidden layers. It follows that the number of units in the input and hidden layer is equal to the number of non-hydrogen atoms of the molecule. The output layer has one unit that computes the molecular property under investigation. The bias unit is connected to the hidden and output units. Using the procedure described above, the network changes the number and significance of the units with each molecule presented to the network.

The bonding relationships between all atoms from the molecular graphs are mapped in the connections between the input and hidden layers. Pairs of atoms exhibiting the same bonding pattern correspond to pairs of units linked by connections with identical weights. The bonding relationship used in the network MolNet-1 considers the type of atoms and bonds on the shortest path between a pair of atoms. Also, a unit that corresponds to an atom i in the input layer is connected to the unit corresponding to the same atom i in the hidden layer by a connection; these connections are classified according to the chemical nature of the atoms. Input-hidden connections corresponding to the same bonding relationship between two atoms either in the same molecule or in different molecules belong to the same class and are characterized by identical weights.

The topology of the connections between the hidden and output layers is determined by the partitioning in classes of the atoms from the molecule presented to MolNet. The atoms are partitioned according to the atomic number Z, the hybridization state and the degree. All atoms from a class, either in the same molecule or in different molecules, correspond to the same type of hidden units. We have to point here that even for units in the same class their contribution to the molecular property depends also on the signal received from the input layer, signal that can be different for units in the same class. Each unit in the hidden layer is connected to the bias neuron by connections partitioned in the same way with the connections between the hidden and output layers, i.e. according to the atom types as defined above. Also, the bias unit is connected with the output unit.

For a molecule with N non-hydrogen atoms, MolNet has N units in the input and hidden layers, respectively. There are N^2 connections between the input and hidden layers, N connections between the hidden and output layers, N connections from the bias unit to the hidden units, and one connection from the bias unit to the output unit. Some connections may have identical weights according to the partitioning schemes described above. Usually, in MolNet the number of adjustable parameters is much smaller than the number of connections.

Each MolNet input unit receives a numerical value representing an invariant computed or determined for the corresponding atom from the molecular graph: the number of hydrogen atoms attached to the atom, the degree, the electronegativity, the atomic charge. Any vertex invariant of the molecular graph can be used as input for MolNet.[28,29]



1

As an example of MolNet topology we present the process of network generation for 1-ethyl-1,2-dimethylcyclopropane **1**. In the case of alkanes the bonding relationship that determines the connection types between the input and hidden layers (IH connections) considers only the topological distance between the carbon atoms. The topological distance between vertices i and j is denoted by d_{ij}, and is equal to the number of bonds on the shortest path between the vertices i and j.[28,29] Distances d_{ij} are elements of the distance matrix of the molecular graph G, $\mathbf{D} = \mathbf{D}(G)$. The distance matrix of the molecular graph of **1**, **D(1)**, computed with the Floyd-Warshall algorithm,[30] is:

D(1)

	1	2	3	4	5	6	7
1	0	1	1	1	1	2	2
2	1	0	1	2	2	3	1
3	1	1	0	2	2	3	2
4	1	2	2	0	2	3	3
5	1	2	2	2	0	1	3
6	2	3	3	3	1	0	4
7	2	1	2	3	3	4	0

The alkane **1** has 7 carbon atoms, giving for the corresponding MolNet network 7 input and hidden units, respectively. As is it apparent from Figure 1a-e each atom from the molecular graph **1** corresponds to a unit with the same label in the input and hidden layers of MolNet. An inspection of the distance matrix of **1** gives a total of 5 classes of topological distances, from 0 to 4. These 5 distance classes correspond to 5 IH connection types (i.e. five parameters that are adjusted in the calibration phase). We have to point here that the bonding relationship between two atoms i and j corresponds to

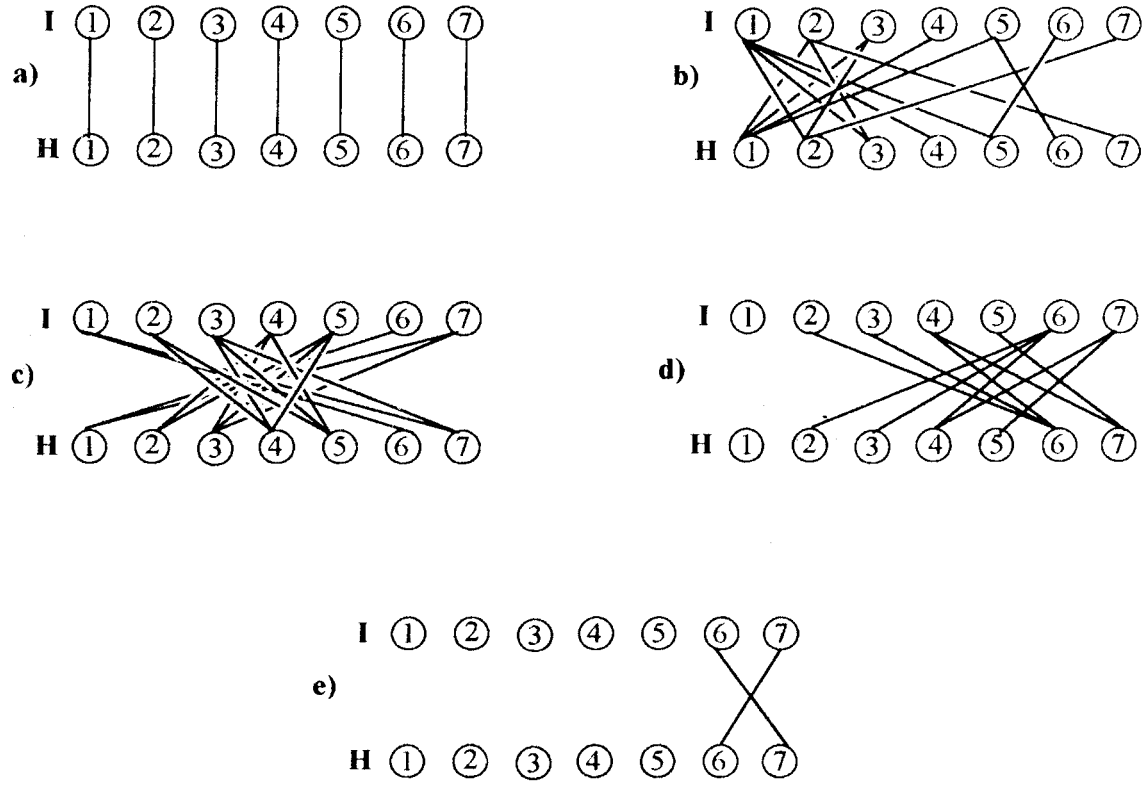


Fig. 1 – The structure of the MolNet connections between the input (**I**) and hidden (**H**) layers for 1-ethyl-1,2-dimethylcyclopropane **1**; each neuron corresponds to the carbon atom with the same label from **1**. The connections between atoms with the same label are presented in a); the connections between atoms situated at distances 1, 2, 3 and 4 are presented in b), c), d) and e), respectively.

2 IH connections with identical weights: one from input unit i to hidden unit j, and one from input unit j to hidden unit i. For example, from **D(1)** one can see that in molecule **1** there are 2 carbon atoms at distance 4, namely 6 and 7. This gives 2 IH connections, depicted in Figure 1e: one connection from input unit 6 to hidden unit 7, and another one from input unit 7 to hidden unit 6. Both connections belong to the same class and they have identical weights.

In Figure 1a-e we present the structure of IH connections according to the classes of identical weights: there are 7 connections corresponding to the distance 0 interactions (Figure 1a), which in our case have identical weights because all non-hydrogen atoms are carbon atoms; the 7 pairs of atoms situated at distance 1 correspond to the 14 connections in Figure 1b; Figure 1c presents the 16 connections between the 8 pairs of atoms situated at distance 2; the 10 connections corresponding to the 5 pairs of atoms situated at distance 3 are presented in Figure 1d; the pair of atoms situated at distance 4 gives 2 IH connections represented in Figure 1e.

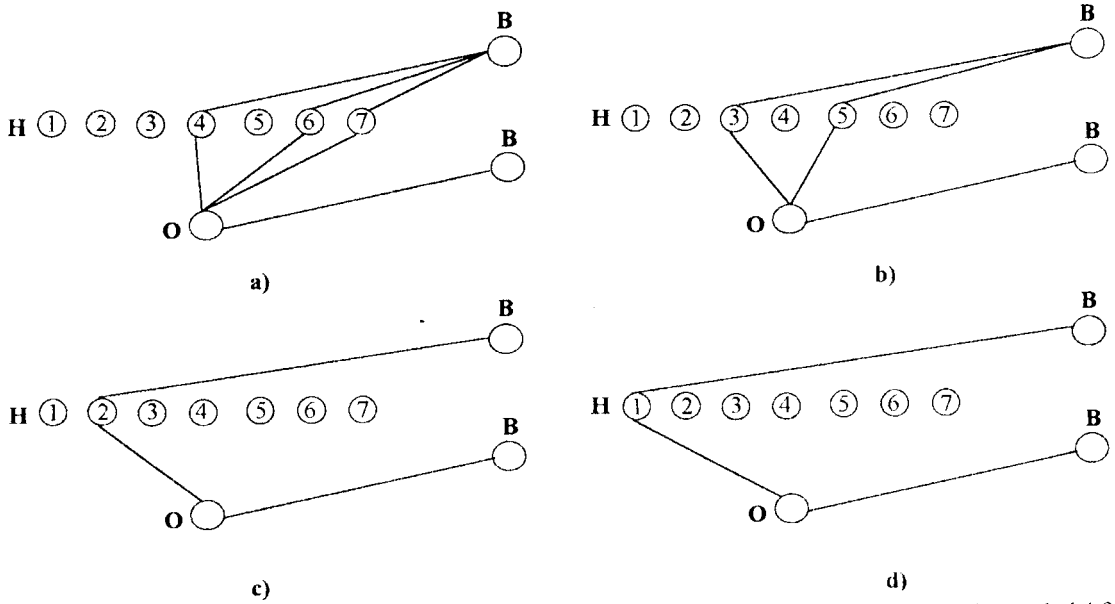


Fig. 2 - The structure of the MolNet connections between the hidden (**H**) and output (**O**) layers for 1-ethyl-1,2-dimethylcyclopropane **1**; the bias neuron is labeled with **B**. The connections to/from atoms with the degree 1, 2, 3 and 4 are presented in a), b), c) and d), respectively.

In alkanes all vertices in the corresponding molecular graph represent sp^3-hybridized carbon atoms. It follows that the connection between a hidden unit and the output unit (a HO connection) is classified according to the degree of the carbon atom represented by the hidden unit. Units representing atoms with identical degrees are linked to the output unit with connections having identical weights. Because the molecular graph of 1-ethyl-1,2-dimethylcyclopropane contains 3 atoms with degrees 1, 2 atoms with degrees 2, one atom with degree 3, and one atom with degree 4, the HO connections belong to four classes (i.e. adjustable weights). The connections between the bias unit and the units in the hidden layer (the BH connections) are classified according to the same rules used for the HO connections, giving in the case of molecule **1** 4 connection classes. For example, molecule **1** contains 2 atoms with degrees 2, namely 3 and 5. The 2 atoms correspond to the hidden units 3 and 5, respectively, and their HO and BH connections are presented in Figure 2b. The 2 HO connections belong to the same class and they have identical weights. The 2 BH connections to units 3 and 5 are identical, too.

The structure of BH and HO connections is presented in Figure 2a-d: the bias and output connections to the units representing the 3 atoms with degrees 1 are presented in Figure 2a; those connecting the atoms with degrees 2, 3, and 4 are depicted in Figure 2b-d. The bias unit has also a connection to the output unit (BO connection). The total number of adjustable parameters for 1-ethyl-1,2-dimethylcyclopropane is: 5 (IH connections) + 4 (BH connections) + 4 (HO connections) + 1 (BO connection) = 14.

MolNet is an MLF network and its use involves two phases: a calibration (learning) and a prediction phase, respectively. The target of the calibration phase is to compute an optimal set of IH, HO, BH and BO connection types (adjustable parameters) that minimizes the error in computing the investigated molecular property for the calibration set of molecules. The selection of the optimization algorithm used to determine an optimal set of connection types depends on the number of input units, on the number of patterns in the calibration set, and on the number of parameters. There is a large variety of methods that can be used, like the global optimization algorithms (simulated annealing or genetic algorithms), the simplex algorithm, direction-set methods (Powell's method), methods that require the computation of the first derivatives like conjugate gradient methods (Fletcher-Reeves or Polak-Ribiere) or quasi-Newton (variable metric) methods (Davidon-Fletcher-Powell or Broyden-Fletcher-Goldfarb-Shanno). In the present investigation we apply the most widely used method in the optimization of neural networks, the backpropagation with momentum algorithm.[2] All MolNet connections are adjusted after the presentation of each molecule. Obviously, if a connection type is absent from a certain molecule its value is not changed after the presentation of that molecule to the network. The connections of the same type from a molecule are adjusted with the same quantity obtained by a summation of individual gradients and application of the usual backpropagation with momentum equation. The learning rate and momentum are set to values that ensure a rapid convergence of the calibration phase.

The prediction power of the MolNet model is tested with a set of molecules that were not used in the calibration phase. In the prediction phase the molecular properties are computed with the weights determined in the calibration phase. If a molecule from the prediction set contains bonding relationships that are absent in the molecules used in the calibration phase these bonding relationships are neglected in predicting the molecular property.

MolNet Operation

Data Set. MolNet is used for the computation of alkane densities at 25°C using as input parameters molecular graph descriptors. The alkanes are separated into a calibration (learning) set and a prediction (test) set. In this way it is possible to determine the MolNet precision in predicting the density for alkanes that are not used in the calibration of the model. The learning set contains 108 alkanes, while the test set contains 25 alkanes between C_6 and C_{10}. The structure and experimental densities of the alkanes used in the present investigation are taken from the literature[9] and are reported in Tables 1 and 2. One alkane used in ref. 9, 2,2,3,3-tetramethylbutane, was deleted from the calibration set because its reported experimental density of 821.70 kg/m^3 is too high when compared with the density of similar alkanes. The separation of alkanes in calibration and prediction sets is identical with that used in ref. 9.

Number of Adjustable Parameters. Because MolNet has a topology that changes with each molecule presented to the network, the number of adjustable parameters (connection types) depends on the structure of the molecules from the learning set. There is one IH connection type corresponding to connections between input unit *i* and hidden unit *i*. In the learning set of 108 alkanes the maximum graph distance between two carbon atoms is 7, giving a total of 8 IH connection classes. The degree of the carbon atoms is from 1 up to 4, giving 4 HO connection types and 4 BH adjustable connections. The total number of adjustable weights for the alkane learning set is: 8 (IH connections) + 4 (BH connections) + 4 (HO connections) + 1 (BO connection) = 17. The ratio between the number of alkanes in the calibration set and the number of adjustable weights is 6.35, a fairly high value indicating that there is no danger of overfitting.

Input Data. The input vectors for the input neurons of MolNet are three molecular graph invariants, namely the degree **DEG**,[28,29] the distance sum **DS**,[31,32] and the reciprocal distance sum **RDS**.[33,34] The degree of an atom *i* from a molecular graph *G* is computed with the equation:

$$\mathbf{DEG}_i = \sum_{j=1}^{N} \mathbf{A}_{ij}$$

Table 1

Alkanes used in MolNet calibration, experimental densities at 25°C (ρ_{exp}), and calibration residuals ($\rho_{exp} - \rho_{cal}$) for MolNet network using DEG (DEG_{res}), and RDS (RDS_{res}) input atomic descriptors

Hydrocarbon	ρ_{exp}	DEG_{res}	RDS_{res}	Hydrocarbon	ρ_{exp}	DEG_{res}	RDS_{res}
3-M-C_5	659.76	−10.81	−9.98	2,5-M_2-C_8	726.40	−0.18	−0.95
2,2-M_2-C_4	644.46	−7.21	−5.76	3,4-M_2-C_8	741.80	−0.52	0.10
2,3-M_2-C_4	657.02	−5.66	−4.99	3,5-M_2-C_8	732.90	−3.13	−2.37
3-M-C_6	682.88	−0.32	−0.25	3,6-M_2-C_8	732.90	−3.27	−3.04
3-E-C_5	693.92	−0.75	1.46	4,4-M_2-C_8	731.20	−5.09	−5.65
2,2-M_2-C_5	669.48	1.73	2.13	4,5-M_2-C_8	743.20	2.52	2.23
2,3-M_2-C_5	690.81	0.44	1.86	4-nP-C_7	732.10	3.10	−1.07
2,4-M_2-C_5	668.23	−2.23	−1.14	4-iP-C_7	735.40	−0.48	−3.51
3,3-M_2-C_5	689.16	1.15	1.69	2-M-3-E-C_7	739.80	−0.15	−0.49
2,2,3-M_3-C_4	685.64	4.62	2.81	2-M-4-E-C_7	732.20	1.95	0.02
n-C_8	698.54	7.10	4.48	3-M-4-E-C_7	746.60	2.09	0.89
2-M-C_7	693.87	6.76	4.09	3-M-5-E-C_7	736.80	−3.08	−3.93
3-M-C_7	701.73	4.96	3.22	2,2,3-M_3-C_7	738.50	−5.67	−4.58
2,4-M_2-C_6	696.17	1.12	0.73	2,2,4-M_3-C_7	725.70	−5.78	−3.23
2,5-M_2-C_6	689.37	9.13	5.13	2,2,5-M_3-C_7	728.10	1.28	0.17
3,3-M_2-C_6	705.95	1.39	1.02	2,2,6-M_3-C_7	723.80	1.58	4.42
3,4-M_2-C_6	715.15	−0.48	1.20	2,3,3-M_3-C_7	748.80	−3.07	−1.95
3-E-2-M-C_5	715.20	0.16	2.30	2,3,4-M_3-C_7	748.50	1.75	0.75
3-E-3-M-C_5	723.54	−0.11	1.70	2,3,5-M_3-C_7	745.10	2.61	3.44
2,2,3-M_3-C_5	712.03	−0.14	−0.33	2,3,6-M_3-C_7	734.70	−0.41	1.12
2,2,4-M_3-C_5	687.84	1.70	3.25	2,4,4-M_3-C_7	734.60	−6.27	−2.24
2,3,3-M_3-C_5	722.30	0.82	0.68	2,4,5-M_3-C_7	737.30	−4.84	−3.16
2,3,4-M_3-C_5	715.09	0.14	2.86	2,4,6-M_3-C_7	719.00	−11.19	−7.25
2-M-C_8	709.60	4.96	1.92	2,5,5-M_3-C_7	736.20	1.72	0.96
3-M-C_8	716.70	2.85	1.07	3,3,5-M_3-C_7	739.00	−7.03	−5.42
3-E-C_7	722.50	3.35	1.55	3,4,4-M_3-C_7	753.50	−6.19	−3.62
4-E-C_7	722.30	5.47	2.70	3,4,5-M_3-C_7	751.90	−4.77	−2.78
2,2-M_2-C_7	706.60	4.81	1.53	2-M-3-iP-C_6	743.60	−3.14	−2.40
2,3-M_2-C_7	722.00	2.39	0.94	3,3-E_2-C_6	757.50	−1.48	−2.19
2,4-M_2-C_7	711.50	−0.80	−1.02	3,4-E_2-C_6	747.20	−3.22	−3.93
2,5-M_2-C_7	713.60	4.16	1.59	2,2-M_2-3-E-C_6	744.70	−3.67	−3.81
2,6-M_2-C_7	704.50	2.13	0.28	2,2-M_2-4-E-C_6	730.20	−2.68	−4.30
3,3-M_2-C_7	721.60	2.42	−0.01	2,3-M_2-3-E-C_6	759.90	−3.33	2.33
3,4-M_2-C_7	727.50	−0.93	−0.79	2,3-M_2-4-E-C_6	751.60	−2.35	−1.37
3,5-M_2-C_7	716.60	−3.69	−4.14	2,4-M_2-4-E-C_6	751.40	0.41	2.47
3-E-3-M-C_6	736.00	−0.19	−0.08	3,3-M_2-4-E-C_6	759.80	5.37	2.97
4-E-2-M-C_6	724.20	7.80	6.16	3,4-M_2-4-E-C_6	759.60	−11.04	−9.42
2,2,4-M_3-C_6	711.80	0.12	−0.23	2,2,3,3-M_4-C_6	760.89	−8.09	−6.35
2,2,5-M_3-C_6	703.20	7.88	2.42	2,2,3,4-M_4-C_6	751.30	−9.42	−6.47
2,3,3-M_3-C_6	733.50	−3.60	−3.98	2,2,3,5-M_4-C_6	733.60	−8.35	−8.35
2,3,4-M_3-C_6	735.10	−2.24	−0.26	2,2,4,5-M_4-C_6	731.61	−4.02	3.80
2,3,5-M_3-C_6	717.90	0.68	−0.31	2,2,5,5-M_4-C_6	714.80	−0.75	5.02
2,4,4-M_3-C_6	720.05	2.65	−0.93	2,3,3,4-M_4-C_6	765.60	−9.30	−6.02
3,3,4-M_3-C_6	741.40	−5.21	−4.03	2,3,3,5-M_4-C_6	744.90	−9.43	5.97
3,3-E_2-C_5	749.92	−1.37	−0.04	2,3,4,4-M_4-C_6	758.60	10.79	6.34
3-E-2,2-M_2-C_5	731.00	−7.14	−5.48	2,3,4,5-M_4-C_6	745.60	11.88	−9.07
3-E-2,3-M_2-C_5	750.80	−2.65	−1.29	3,3,4,4-M_4-C_6	778.90	−4.96	−2.24
2,2,3,3-M_4-C_5	752.97	−1.81	−3.28	2,4-M_2-3-iP-C_5	754.57	−6.42	−0.18
2,2,3,4-M_4-C_5	735.22	5.66	−2.24	2-M-3,3-E_2-C_5	775.50	0.41	1.20
2,3,3,4-M_4-C_5	751.11	−4.49	−2.88	2,2,3-M_3-3-E-C_5	778.00	−4.58	−2.31
3-E-C_8	735.90	1.40	0.51	2,2,4-M_3-3-E-C_5	753.10	−10.18	−4.61
4-E-C_8	734.30	3.63	0.27	2,3,4-M_3-3-E-C_5	773.50	−6.61	−3.71
2,2-M_2-C_8	720.80	1.56	0.89	2,2,3,3,4-M_5-C_5	776.75	−10.85	−6.45
2,4-M_2-C_8	722.60	−5.69	−5.12	2,2,3,4,4-M_5-C_5	763.61	6.79	0.59

Table 2

Alkanes used in MolNet prediction, experimental densities at 25°C (ρ_{exp}), and prediction residuals ($\rho_{exp} - \rho_{cal}$) for MolNet network using **DEG** ($\mathbf{DEG}_{res}$), and **RDS** ($\mathbf{RDS}_{res}$) input atomic descriptors

Hydrocarbon	ρ_{exp}	$\mathbf{DEG}_{res}$	$\mathbf{RDS}_{res}$	Hydrocarbon	ρ_{exp}	$\mathbf{DEG}_{res}$	$\mathbf{RDS}_{res}$
3-E-2,4-M_2-C_5	734.10	-4.04	-0.39	2,3-M_2-C_6	708.16	4.55	3.70
2-M-C_5	648.52	-11.47	-10.27	3-M-3-E-C_7	746.30	2.53	-2.72
2,2,4,4-M_4-C_5	715.61	5.02	9.58	4-M-C_8	716.30	4.17	1.55
n-C_7	679.50	2.72	1.77	4-M-3-E-C_7	746.80	0.03	0.38
2,3-M_2-C_8	734.40	2.65	-1.34	4,4-M_2-C_7	718.30	3.76	-3.79
2-M-C_6	674.34	3.80	2.39	4-M-4-E-C_7	747.20	-1.28	-1.19
2,6-M_2-C_8	723.60	-6.20	-4.51	3-E-2-M-C_6	729.00	3.81	2.76
4-M-C_7	700.71	2.86	2.04	3,3,4-M_3-C_7	752.70	-5.22	3.99
2,7-M_2-C_8	720.20	-1.54	0.26	3-E-4-M-C_6	738.00	2.49	3.53
3-E-C_6	709.45	4.26	3.94	2,5-M_2-3-E-C_6	736.80	-0.74	-1.64
3,3-M_2-C_8	735.10	-3.02	-1.98	2,2,3-M_3-C_6	725.70	0.44	-2.74
2,2-M_2-C_6	691.11	10.20	6.35	2,2,4,4-M_4-C_6	742.40	-2.49	1.85
2-M-5-E-C_7	731.80	-0.90	1.58				

where $\mathbf{A} = \mathbf{A}(G)$ is the adjacency matrix of G. The degree vector of 1-ethyl-1,2-dimethylcyclopropane is $\mathbf{DEG(1)} = \{4, 3, 2, 1, 2, 1, 1\}$. This graph invariant is not a very discriminat one, and in the above example the non-equivalent atoms 4, 6, and 7 have identical degrees.

In a molecular graph G the distance sum of an atom i is the sum of the elements in the ith row (or ith column) of the distance matrix $\mathbf{D} = \mathbf{D}(G)$:

$$\mathbf{DS}_i = \sum_{j=1}^{N} \mathbf{D}_{ij}$$

The distance sum vector of 1-ethyl-1,2-dimethylcyclopropane is $\mathbf{DS(1)} = \{8, 10, 11, 13, 11, 16, 15\}$. Although it is much more discriminant than the degree, the distance sum is degenerated for the non-equivalent atoms 3 and 5 from **1**.

In a molecular graph G the reciprocal distance sum of the atom i is defined by the equation:

$$\mathbf{RDS}_i = \sum_{j=1}^{N} \mathbf{RD}_{ij}$$

where $\mathbf{RD} = \mathbf{RD}(G)$ is the reciprocal distance matrix of G. The reciprocal distance sum vector of 1-ethyl-1,2-dimethylcyclopropane is $\mathbf{RDS(1)} = \{5.00000, 4.33333, 3.83333, 3.16667, 3.83333, 2.75000, 2.91667\}$. In this example, **RDS** is degenerated for the atoms 3 and 5.

Learning Method. In the calibration phase the MolNet weights are optimized with the standard backpropagation with momentum method[2] until the convergence is obtained. The calibration phase converges when the correlation coefficient between experimental and calculated alkane density values improves by less than 10^{-5} in 100 epochs. One epoch corresponds to one complete presentation of the 108 molecules in the learning set. The connection weights are updated after the presentation of each molecule. Random values between –0.1 and 0.1 are used as initial weights. The calibration process is very sensitive to the learning rate and momentum values, and the results from several trials with different values for the momentum are presented in this study. The learning rate values for both the hidden and output layers are equal to 0.05. The momentum is set between 0.30 and 0.05 for all activation functions used in this study. Both learning rate and momentum values are maintained constant during the training phase. In all cases the calibration phase converges after a few hundreds epochs and the results are slightly influenced by the initial random set of weights.

Activation Functions. The sigmoid, which is the most commonly used activation function, takes values between 0 and 1. For large negative arguments its value is close to 0, and practice demonstrated that learning with the backpropagation algorithm is difficult in such conditions. To overcome this deficiency of the sigmoid function, the hyperbolic tangent (tanh) which takes values between –1 and 1 is used in the present study. Because the tanh and the sigmoid activation functions are very flat when the absolute value of the argument is greater than 10 their derivatives have small values, leading to a poor sensitivity of the two activation functions to large positive or negative arguments. This behavior represents an important cause of the very slow rates of convergence during the training of neural networks with algorithms that use the derivative of the activation function, like the backpropagation algorithm. A linear output activation function overcomes the problems of the sigmoidal function because it has a constant sensitivity. For the output layer we use also a linear activation function. A new type of activation function is the symmetric logarithmoid,[35,36] defined by the formula: Act (z) = sign (z) ln (1 + $|z|$). The symmetric logarithmoid (symlog) is a monotonically increasing function with the maximum sensitivity near zero and with a monotonically decreasing sensitivity away from zero. Because its output is not restricted to a finite range of values this function is sensitive to large positive or negative arguments. The symlog function is used for the output layer. The linear and symlog functions are unbounded and this property makes impossible their use in the hidden layers.

Preprocessing of the Data. Each input vector (**DEG**, **DS**, or **RDS**) and output values (representing the target alkane densities) are linearly scaled between –0.9 and 0.9. We have to point out that for the tanh output activation function the scaling is required by the range of values of the function between –1 and 1, while for the unbounded functions (linear and symlog) the experience shows that a linear scaling improves the calibration process.

Performance Indicators. The MolNet performances are evaluated both for the network calibration and prediction. The quality of MolNet calibration is estimated by comparing the calculated alkane densities at the end of the calibration process (ρ_{cal}) with the target values (ρ_{exp}), while the predictive quality is estimated with a set of alkanes that were not used in the calibration phase by comparing the predicted (ρ_{pr}) and experimental densities. The performances of different MolNet networks are compared by using the correlation coefficient r and the standard deviation s of the linear correlation between experimental and calculated (in calibration or prediction) densities: $\rho_{exp} = A + B \cdot \rho_{cal/pr}$.

MolNet Computation of Alkane Densities

Usually, the number of hidden layers and the number of units in each hidden layer of a MLF neural network is determined during the calibration phase. As presented above, MolNet has a topology that depends on the structure of each molecule presented to the network. Therefore, MolNet has always one hidden layer and the optimization of the number of hidden units is no longer required because it is determined by the number of atoms in each molecule. MolNet accepts as input any atomic property computed on the basis of the structure of the molecular graph. In previous MolNet applications in QSPR studies we have noticed that the calibration and prediction results depend on the input atomic invariant. This study investigates three input atomic descriptors, namely **DEG**, **DS**, and **RDS**. These descriptors can be easily computed for any molecule from the structure of the corresponding molecular graph.

In the first set of experiments the MolNet parameters are optimized using as input data the **DEG** atomic descriptor. The calibration and prediction results are reported in Table 3. The calibration correlation coefficient, r_{cal}, is in the range 0.985 to 0.989, and the calibration standard deviation, s_{cal}, takes values between 3.95 kg/m^3 and 4.61 kg/m^3. The prediction correlation coefficient, r_{pr}, takes values between 0.976 and 0.985, while the prediction standard deviation, s_{pr}, is in the range 4.43 kg/m^3 to 5.68 kg/m^3. Overall, the best calibration and prediction results are obtained with linear output function, followed by the symlog output function. The alkane density residuals computed with **DEG** input data, a linear output function, a hidden and an output momentum $\alpha_h = \alpha_o = 0.05$ are presented in column 4 of

Table 3

MolNet calibration and prediction results for the computation of alkane densities using **DEG** input atomic descriptor. The table reports the number of training epochs, the hidden and output momentum (α_h and α_o), the output activation function, the calibration and prediction standard deviation (s_{cal} and s_{pr}) and correlation coefficient (r_{cal} and r_{pr}). All networks were provided with the tanh hidden activation function.

Epoch	Hidden Momentum (α_h)	Output Activation Function	Output Momentum (α_o)	s_{cal}	r_{cal}	s_{pr}	r_{pr}
1000	0.30	linear	0.30	4.13	0.988	4.81	0.983
1500	0.30	linear	0.15	4.00	0.989	4.75	0.983
400	0.30	linear	0.10	3.95	0.989	4.56	0.985
500	0.30	linear	0.05	3.94	0.989	4.56	0.985
1600	0.15	linear	0.05	3.95	0.989	4.47	0.985
1300	0.10	linear	0.05	3.97	0.989	4.47	0.985
1200	0.05	linear	0.05	3.99	0.989	4.43	0.985
600	0.15	linear	0.15	4.02	0.988	4.52	0.985
1000	0.15	linear	0.10	3.99	0.989	4.51	0.985
700	0.10	linear	0.10	4.00	0.989	4.47	0.985
1500	0.30	symlog	0.30	4.15	0.988	5.56	0.977
1900	0.30	symlog	0.15	4.08	0.988	5.50	0.977
1800	0.30	symlog	0.10	4.08	0.988	5.45	0.978
1900	0.30	symlog	0.05	4.08	0.988	5.43	0.978
2000	0.15	symlog	0.05	4.03	0.988	5.31	0.979
1900	0.10	symlog	0.05	4.03	0.988	5.26	0.979
1800	0.05	symlog	0.05	4.01	0.989	5.24	0.980
2000	0.15	symlog	0.15	4.05	0.988	5.32	0.979
1700	0.15	symlog	0.10	4.03	0.988	5.33	0.979
1900	0.10	symlog	0.10	4.03	0.988	5.27	0.979
500	0.30	tanh	0.30	4.61	0.985	5.68	0.976
1500	0.30	tanh	0.15	4.54	0.985	5.58	0.977
1700	0.30	tanh	0.10	4.53	0.985	5.57	0.977
1500	0.30	tanh	0.05	4.53	0.985	5.56	0.977
1100	0.15	tanh	0.05	4.46	0.986	5.19	0.980
800	0.10	tanh	0.05	4.50	0.986	5.42	0.978
1400	0.05	tanh	0.05	4.36	0.986	5.12	0.980
800	0.15	tanh	0.15	4.51	0.985	5.45	0.978
900	0.15	tanh	0.10	4.51	0.985	5.44	0.978
800	0.10	tanh	0.10	4.50	0.986	5.43	0.978

Table 1 for calibration, and in Table 2 for prediction. This example, selected because it offers good calibration and prediction results, has the following statistical indices: $r_{cal} = 0.989$, $s_{cal} = 3.99$ kg/m^3, $r_{pr} = 0.985$, and $s_{pr} = 4.43$ kg/m^3. Usually, in a QSPR model an outlier is defined as the pattern with an absolute residual 3 times greater than the standard deviation. In this case, when $3s_{cal} = 11.97$ and $3s_{pr} = 13.29$, there is no calibration or prediction outlier. From the whole set of 108 alkanes used in calibration, one finds 13 cases with absolute residuals between $2s_{cal}$ and $3s_{cal}$. The set of 13 alkanes is presented here together with their corresponding residuals: 3-methylpentane with −10.81, 2,5-dimethylhexane with 9.13, 2,4,6-trimethylheptane with −11.19, 3,4-dimethyl-4-ethylhexane with −11.04,

2,2,3,3-tetramethylhexane with –8.09, 2,2,3,4-tetramethylhexane with –9.42, 2,2,3,5-tetramethylhexane with –8.35, 2,3,3,4-tetramethylhexane with –9.30, 2,3,3,5-tetramethylhexane with –9.43, 2,3,4,4-tetramethylhexane with –10.79, 2,3,4,5-tetramethylhexane with –11.88, 2,2,4-trimethyl-3-ethylpentane with –10.18, 2,2,3,3,4-pentamethylpentane with –10.85. From the above list one can see that larger errors are obtained for higher alkanes when **DEG** is used as input data. The prediction residuals are small, with only 2 alkanes presenting absolute residuals between $2s_{pr}$ and $3s_{pr}$: 2-methylpentane with –11.47, and 2,2-dimethylhexane with 10.20.

The second test considers the **DS** atomic descriptor as input data. The MolNet calibration and prediction results are presented in Table 4. In the calibration phase r_{cal} is between 0.950 and 0.974, and

Table 4

MolNet calibration and prediction results for the computation of alkane densities using **DS** input atomic descriptor. The notations are explained in Table 3.

Epoch	Hidden Momentum (α_h)	Output Activation Function	Output Momentum (α_o)	s_{cal}	r_{cal}	s_{pr}	r_{pr}
400	0.30	linear	0.30	7.06	0.964	7.78	0.954
1700	0.30	linear	0.15	6.65	0.968	9.29	0.934
2000	0.30	linear	0.10	6.12	0.973	8.71	0.942
1900	0.30	linear	0.05	6.10	0.973	8.62	0.943
1700	0.15	linear	0.05	6.04	0.974	7.74	0.955
1900	0.10	linear	0.05	6.05	0.974	7.45	0.958
1800	0.05	linear	0.05	6.07	0.974	7.17	0.961
600	0.15	linear	0.15	6.81	0.967	7.28	0.960
600	0.15	linear	0.10	6.77	0.967	7.23	0.961
2000	0.10	linear	0.10	6.08	0.973	7.48	0.958
1000	0.30	symlog	0.30	7.18	0.963	9.14	0.936
900	0.30	symlog	0.15	6.77	0.967	7.76	0.954
1900	0.30	symlog	0.10	6.47	0.970	7.42	0.958
1800	0.30	symlog	0.05	6.49	0.970	7.37	0.959
1600	0.15	symlog	0.05	6.51	0.969	7.75	0.955
1400	0.10	symlog	0.05	7.02	0.965	9.02	0.938
1200	0.05	symlog	0.05	6.75	0.967	6.87	0.965
1400	0.15	symlog	0.15	6.95	0.965	6.44	0.969
1600	0.15	symlog	0.10	6.51	0.969	7.64	0.956
1300	0.10	symlog	0.10	6.60	0.969	7.52	0.957
2000	0.30	tanh	0.30	7.22	0.962	8.25	0.948
1900	0.30	tanh	0.15	6.99	0.965	8.48	0.945
1700	0.30	tanh	0.10	8.33	0.950	8.15	0.950
1900	0.30	tanh	0.05	6.87	0.966	8.62	0.943
1800	0.15	tanh	0.05	6.97	0.965	6.96	0.964
400	0.10	tanh	0.05	6.87	0.966	7.20	0.961
1000	0.05	tanh	0.05	6.98	0.965	6.52	0.968
900	0.15	tanh	0.15	7.39	0.961	9.20	0.935
2000	0.15	tanh	0.10	7.02	0.964	6.85	0.965
1000	0.10	tanh	0.10	7.04	0.964	6.68	0.966

s_{cal} takes values between 6.04 kg/m^3 and 8.33 kg/m^3. In the prediction phase r_{pr} takes values between 0.934 and 0.969, and s_{pr} is in the range 6.44 kg/m3 to 9.29 kg/m^3. Because the models computed with **DS** input data are of lower statistical quality than those obtained with **DEG** input data, we conclude that this descriptor is not of great value in MolNet QSPR studies of alkane densities.

The third test uses the **RDS** atomic descriptor as input data to optimize the MolNet parameters, and Table 5 presents the calibration and prediction results obtained. In the calibration phase r_{cal} is between 0.986 and 0.993, and s_{cal} takes values between 3.10 kg/m^{3}and 4.50 kg/m^3. In the prediction phase r_{pr} takes values between 0.974 and 0.988, and s_{pr} is in the range 4.02 kg/m^3 to 5.88 kg/m^3. The statistical indices obtained with the three output activation functions show that the linear output func-

Table 5

MolNet calibration and prediction results for the computation of alkane densities using **RDS** input atomic descriptor. The notations are explained in Table 3.

Epoch	Hidden Momentum (α_h)	Output Activation Function	Output Momentum (α_o)	s_{cal}	r_{cal}	s_{pr}	r_{pr}
2000	0.30	linear	0.30	3.26	0.992	4.22	0.987
800	0.30	linear	0.15	3.21	0.993	4.19	0.987
300	0.30	linear	0.10	3.25	0.993	4.23	0.987
1900	0.30	linear	0.05	3.18	0.993	4.21	0.987
300	0.15	linear	0.05	3.17	0.993	4.12	0.987
1000	0.10	linear	0.05	3.12	0.993	4.06	0.988
1900	0.05	linear	0.05	3.10	0.993	4.03	0.988
1200	0.15	linear	0.15	3.16	0.993	4.10	0.988
900	0.15	linear	0.10	3.15	0.993	4.09	0.988
800	0.10	linear	0.10	3.13	0.993	4.02	0.988
500	0.30	symlog	0.30	4.33	0.987	5.87	0.974
1100	0.30	symlog	0.15	4.27	0.987	5.86	0.974
1100	0.30	symlog	0.10	4.27	0.987	5.88	0.974
1200	0.30	symlog	0.05	4.24	0.987	5.88	0.974
1200	0.15	symlog	0.05	4.19	0.987	5.82	0.975
1700	0.10	symlog	0.05	4.17	0.988	5.81	0.975
1100	0.05	symlog	0.05	4.19	0.987	5.81	0.975
1200	0.15	symlog	0.15	4.21	0.987	5.80	0.975
1200	0.15	symlog	0.10	4.21	0.987	5.82	0.975
1300	0.10	symlog	0.10	4.18	0.988	5.79	0.975
900	0.30	tanh	0.30	4.50	0.986	5.40	0.978
900	0.30	tanh	0.15	4.46	0.986	5.45	0.978
1100	0.30	tanh	0.10	4.46	0.986	5.48	0.978
1000	0.30	tanh	0.05	4.42	0.986	5.46	0.978
1100	0.15	tanh	0.05	4.38	0.986	5.43	0.978
1100	0.10	tanh	0.05	4.37	0.986	5.43	0.978
1100	0.05	tanh	0.05	4.35	0.987	5.42	0.978
1400	0.15	tanh	0.15	4.41	0.986	5.41	0.978
1200	0.15	tanh	0.10	4.40	0.986	5.43	0.978
1100	0.10	tanh	0.10	4.38	0.986	5.46	0.978

tion gives the best results. The prediction results are good, with statistical indices slightly lower than those obtained for calibration. The density residuals computed with **RDS** input data, a linear output function, a hidden momentum and an output momentum $\alpha_h = \alpha_o = 0.05$ are presented in Tables 1 and 2, column 6. This example, with $r_{cal} = 0.993$, $s_{cal} = 3.10$ kg/m^3, $r_{pr} = 0.988$, and $s_{pr} = 4.03$ kg/m^3, is selected because it offers good calibration and prediction results. There are 2 statistical outliers with an absolute residual greater than $3s_{cal}$, presented here with their residuals: 3-methylpentane with –9.98, and 3,4-dimethyl-4-ethylhexane with –9.42. In the calibration set of 108 alkanes there are 7 alkanes with absolute residuals between $2s_{cal}$ and $3s_{cal}$: 2,4,6-trimethylheptane with –7.25, 2,2,3,3-tetramethylhexane with –6.35, 2,2,3,4-tetramethylhexane with –6.47, 2,2,3,5-tetramethylhexane with –8.35, 2,3,4,4-tetramethylhexane with –6.34, 2,3,4,5-tetramethylhexane with –9.07, 2,2,3,3,4-pentamethylpentane with –6.45. The prediction results show that there are two alkanes with absolute residuals between $2s_{pr}$ and $3s_{pr}$: 2-methylpentane with –10.27, and 2,2,4,4-tetramethylpentane with 9.58.

CONCLUSIONS

We have presented a MolNet application for the computation of the alkanes densities by using as input data three structural descriptors derived from the molecular graph, namely the degree, the distance sum, and the reciprocal distance sum. The good calibration and prediction results show that MolNet is an effective way to establish structure-property models. Three output activation functions were used, namely the hyperbolic tangent, the symmetric logarithmoid, and the linear function. The calibration and prediction results clearly show that the output linear function gives the best results. This conclusion is in line with our previous results.[26,27,37]

Each MolNet input unit receives a numerical value representing an invariant computed for the corresponding atom from the molecular graph. The MolNet calibration and prediction results depend strongly on the input structural invariant, and our study demonstrates that the reciprocal distance sum gives the best estimations of the alkane densities. The atomic degree gives also good results, while for distance sum input data the estimations are affected by larger errors. This finding shows that simple to compute atomic invariants are very effective in correlating the alkane densities. The numerical characterization of the molecular structure with the aid of molecular graph descriptors represents an efficient and simple way of generating QSPR models especially in the case of acyclic, highly flexible molecules, as those investigated in this paper. The geometry of acyclic molecules presents many local minima, making thus difficult to characterize in a simple and unique way the local and global structure with geometric or quantum structural descriptors. Only techniques of exploring the conformational space and molecular dynamics simulations can offer information on the population of the low energy conformations. The identification of all conformations and the computation of the molecular population of each minima is computationally expensive, and for these reasons molecular dynamics simulations are not routinely used in QSPR studies.

The fact that **RDS** is the best input atomic invariant is sustained by our earlier investigations.[26,27,37] The **DEG** descriptor considers only the number of neighbors of an atom in the molecular graph, while **RDS** is more complex, reflecting the influence of atoms situated at greater topological distance. The local structure of the molecular graph is reflected in different ways by the two invariants, namely **DEG** and **RDS**. An inspection of the errors shows that there are a number of alkanes with large residuals obtained with both **DEG** and **RDS**. In calibration these alkanes are 3-methylpentane, 2,4,6-trimethylpentane, 3,4-dimethyl-4-ethylhexane, 2,2,3,3-tetramethylhexane, 2,2,3,4-tetramethylhexane, 2,2,3,5-tetramethylhexane, 2,3,4,4-tetramethylhexane, 2,3,4,5-tetramethylhexane, and 2,2,3,3,4-pentamethylpentane. In prediction only 2-methylpentane has large errors computed with both **DEG** and **RDS** input descriptors. There are two explications for this situation: either the experimental values are affected by large errors, or both descriptors are not able to reflect some important structural features. Because **DEG** and **RDS** reflect in a different way the local structure of the molecular graph, it is more probable that the density of the above alkanes is affected by large errors. It is difficult to find a general

structural feature for the alkanes from the above list, because there are similar compounds whose densities are computed with high precision. This is a second argument that suggests that the density of the above alkanes may be in error. Other atomic descriptors computed from the molecular graph have to be investigated with the aim to identify the best MolNet input data for the prediction of density.

REFERENCES

1. Part 10: O. Ivanciuc, *Rev. Roum. Chim.*, **1999**, *44*, 299–309.
2. D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Nature*, **1986**, *323*, 533–536.
3. P. D. Wasserman, "Neural Computing", Van Nostrand Reinhold, New York, 1989.
4. J. Zupan and J. Gasteiger, "Neural Networks for Chemists", VCH, Weinheim, 1993.
5. A. B. Bulsari (Ed.), "Neural Networks for Chemical Engineers", Elsevier, Amsterdam, 1995.
6. J. Devillers (Ed.), "Neural Networks in QSAR and Drug Design", Academic Press, London, 1996.
7. D. W. Elrod, G. M. Maggiora, and R. G. Trenary, *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 477–484.
8. D. W. Elrod, G. M. Maggiora, and R. G. Trenary, *Tetrahedron Comput. Methodol.*, **1990**, *3*, 163–174.
9. A. A. Gakh, E.G. Gakh, B.G. Sumpter, and D.W. Noid, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 832–839.
10. O. Ivanciuc, J.-P. Rabine, and D. Cabrol-Bass, *Comput. Chem.*, **1997**, *21*, 437–443.
11. A. T. Balaban, S. C. Basak, T. Colburn, and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1118–1121.
12. D. Cherqaoui and D. Villemin, *J. Chem. Soc. Faraday Trans.*, **1994**, *90*, 97–102.
13. D. Cherqaoui, D. Villemin, A. Mesbah, J.-M. Cense, and V. Kvasnika, *J. Chem. Soc. Faraday Trans.*, **1994**, *90*, 2015–2019.
14. F. R. Burden, *Quant. Struct.-Act. Relat.*, **1996**, *15*, 7–11.
15. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye, and J. P. Doucet, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 644–653.
16. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye, and J. P. Doucet, *J. Chem. Inf. Comput Sci.*, **1997**, *37*, 587–598.
17. J. H. Schuur, P. Selzer, and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 334–344.
18. J. Gasteiger, J. Sadowski, J. Schuur, P. Selzer, L. Steinhauer, and V. Steinhauer, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 1030–1037.
19. L. H. Hall and C. T. Story, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 1004–1014.
20. S. Hatrik and P. Zahradnik, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 992–995.
21. H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 1205–1213.
22. S.-S. So and M. Karplus, *J. Med. Chem.*, **1997**, *40*, 4347–4359.
23. S.-S. So and M. Karplus, *J. Med. Chem.*, **1997**, *40*, 4360–4371.
24. D. B. Kireev, *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 175–180.
25. I. I. Baskin, V. A. Palyulin, and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 715–721.
26. O. Ivanciuc, *Anal. Chim. Acta*, **1999**, *384*, 271–284.
27. O. Ivanciuc, *Rev. Roum. Chim.*, **1998**, *43*, 885–894.
28. M. V. Diudea and O. Ivanciuc, "Molecular Topology", Comprex, Cluj, Romania, 1995.
29. O. Ivanciuc and A. T. Balaban. *Graph Theory in Chemistry*. In: "The Encyclopedia of Computational Chemistry", Eds.: P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, and P. R. Schreiner. John Wiley & Sons, Chichester, 1998, pp. 1169–1190.
30. B. Mohar and T. Pisanski, *J. Math. Chem.*, **1988**, *2*, 267–277.
31. A.T. Balaban, *Chem. Phys. Lett.*, **1982**, *89*, 399–404.
32. A.T. Balaban, *Pure Appl. Chem.*, **1983**, *55*, 199–206.
33. O. Ivanciuc, *Rev. Roum. Chim.*, **1989**, *34*, 1361–1368.
34. O. Ivanciuc, T.-S. Balaban, and A.T. Balaban, *J. Math. Chem.*, **1993**, *12*, 309–318.
35. A. B. Bulsari and H. Saxén, *Neurocomputing*, **1991**, *3*, 125–133.
36. A. B. Bulsari and H. Saxén, *Neural Network World*, **1991**, *4*, 221–224.
37. O. Ivanciuc, Molecular Graph Descriptors Used in Neural Network Models. In: *Topological Indices and Related Descriptors in QSAR and QSPR*, Eds.: J. Devillers and A. T. Balaban. Gordon and Breach Science Publishers, The Netherlands, 1999, pp. 697–777.