Revue Roumaine de Chimie, **1998**, *43* (9), 885–894

ARTIFICIAL NEURAL NETWORKS APPLICATIONS. PART 9.[1] MOLNET PREDICTION OF ALKANE BOILING POINTS

Ovidiu IVANCIUC

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

Received March 14, 1997

A new neural network, MolNet, is introduced for the computation of molecular properties. MolNet changes its structure (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. Each non-hydrogen atom in the molecule has a corresponding unit in the input and hidden layers, while the output layer has only one unit which gives the computed molecular property. The connections between the input and hidden layers correspond to the bonding relationships of the atoms, with identically weighted connections for pairs of atoms exhibiting the same bonding pattern. The bonding relationship considers the type of atoms and bonds on the shortest path between a pair of atoms. The connections between the hidden and output layers are classified according to the partitioning of the atoms by their atomic number Z, the hybridization state and the degree. The input to the first layer of neurons represents atomic descriptors such as the degree, distance sum, and reciprocal distance sum. MolNet is applied for the computation of boiling points of alkanes, giving good results both in calibration and prediction.

INTRODUCTION

The growing interest in the application of Artificial Neural Networks (ANN)[2–5] in the field of Quantitative Structure-Property Relationships (QSPR) is a result of their demonstrated superiority over the traditional models.[6–14] An important problem for the chemical applications of neural networks remains the numerical representation of the chemical structure. Various structural representations of organic compounds were used in recent QSPR studies using Multi-Layer Feedforward (MLF) neural models: connection table describing the substituents;[6] modified bond-electron matrix containing as structural information the formal bond order between a pair of atoms and the atomic number Z;[7] topological distance;[8] constitutional descriptors and topological indices;[9] numerical code;[10] molecular subgraphs (clusters);[11] vectorial representation of the chemical structure of the substituents;[12] topo-stereochemical code describing the environment of an atom,[13,14] molecular graph distance counts.[15]

Usually, the MLF networks used in QSPR studies receive information regarding the molecular structure of the chemical compounds only through the agency of the input neurons that receive a numerical representation of certain structural descriptors. Therefore, 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 proposed: the ChemNet introduced by Kireev,[16] the Baskin, Palyulin, Zefirov (BPZ) neural device,[17] and MolNet defined by Ivanciuc.[18] 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, and accept a wide variety of atomic indices as input structural descriptors. We present here an application of MolNet to the computation of boiling points of alkanes.

MolNet description

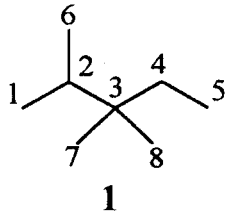
MolNet is a new type of neural network that changes the topology according to the structure of each molecule presented to the network. Each non-hydrogen atom in the molecule has a corresponding unit in the input and hidden layers, while the output layer has only one unit, corresponding to the molecular property under investigation. The network is provided with a bias unit, connected to the hidden and output units. Therefore, the network changes the number and significance of the units with each molecule presented to the network. The connections between the input and hidden layers correspond to the bonding relationships of the atoms, with identically weighted connections for pairs of atoms exhibiting the same bonding pattern. For the network MolNet-1 the bonding relationship 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 have identical weights.

The connections between the hidden and output layers are classified according to the partitioning of the atoms by their atomic number Z, the hybridization state and the degree. This partitioning scheme of the connections defines an atom-type contribution to the molecular property that is investigated. We have to point here that even for atoms 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 atoms in the same class. The bias neuron is connected to each neuron in the hidden layer 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 neuron is connected with the output neuron. For a molecule with N non-hydrogen atoms, there are N^2 connections between the input and hidden layers, N connections between the hidden and output layers, and N + 1 connections from the bias neuron. Some connections may have identical weights according to the partitioning schemes described above.

The data sent to the N input units is a vector containing a property computed at the atomic level: the number of hydrogen atoms attached to each atom, the degree, the electronegativity, the atomic charge. Any vertex invariant of the molecular graph[19,20] can be used as input for MolNet.

For alkanes the bonding relationship that determines the connection types between the input and hidden layers (IH connections) takes into account only the topological distance between the carbon atoms. As an example of MolNet topology we will present the structure of the network for 2,3,3-trimethylpentane (**1**); the molecular graph of 2,3,3-trimethylpentane is presented in Fig. 1. 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. Distances d_{ij} are elements of the distance matrix of the molecular graph G, **D**(G).[19,20] The distance matrix of the molecular graph of **1**, **D**(**1**), computed with the Floyd-Warshall algorithm,[21] is presented in Table 1.

Each atom from the molecular graph **1** corresponds to a unit with the same label in the input and hidden layers of MolNet, as presented in Fig. 2a-e. As is apparent from the distance matrix of **1**,



1

Fig. 1. – The molecular graph of 2,3,3-trimethylpentane **1** (2,3,3-M_3-C_5).

Table 1

The distance matrix of the molecular graph of 2,3,3-trimethylpentane **1**

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0	1	2	3	4	2	3	3
2	1	0	1	2	3	1	2	2
3	2	1	0	1	2	2	1	1
4	3	2	1	0	1	3	2	2
5	4	3	2	1	0	4	3	3
6	2	1	2	3	4	0	3	3
7	3	2	1	2	3	3	0	2
8	3	2	1	2	3	3	2	0

five classes of topological distances, from 0 to 4, are found in this molecule, corresponding to five IH connection types (i.e. five parameters that are adjusted in the learning phase). In Fig. 2a-e we present the structure of IH connections according to the classes of identical weights: there are 8 connections corresponding to the distance 0 (Fig. 2a), which in our case have identical weights because all non-

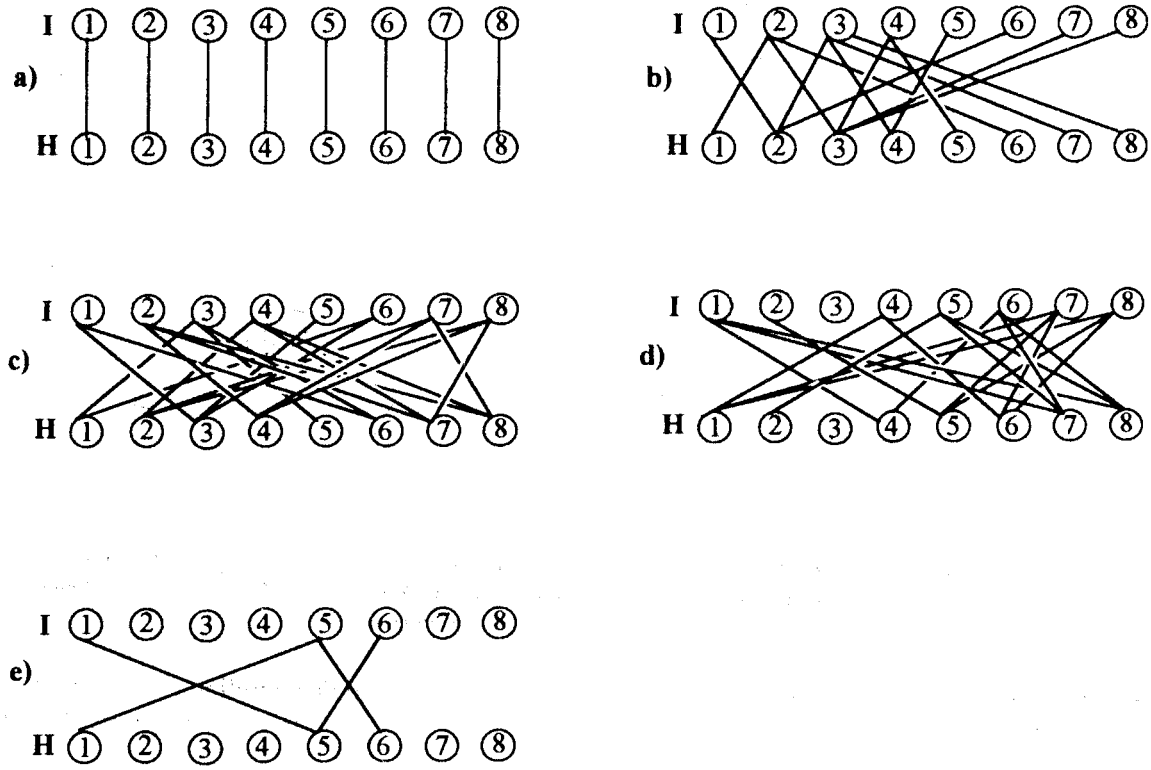


Fig. 2. – The structure of the MolNet connections between the input (**I**) and hidden (**H**) layers for 2,3,3-trimethylpentane; each neuron corresponds to the carbon atom with the same label from Figure 1. The connections between atoms with the same label are presented in a); the connections between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

hydrogen atoms are carbon atoms; the 14 connections in Fig. 2b are connecting atoms situated at distance 1; Fig. 2c presents the 20 connections between atoms situated at distance 2; the 18 connections corresponding to distance 3 are presented in Fig. 2d; Fig. 2e depicts the 4 connections representing relations between atoms situated at distance 4.

In alkanes the connections between the hidden and output layers (the HO connections) are classified according to the degree of the carbon atoms: atoms with identical degree are linked to the output unit with connections having identical weights. Because the molecular graph of 2,3,3-trimethylpentane contains 5 atoms with degree 1, and one atom with degree 2, 3, and 4, respectively, 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 four adjustable weights. The structure of BH and HO connections is presented in Fig. 3a-d: the bias and output connections to atoms with the degree equal to 1 are presented in Fig. 3a; those connecting the atoms with degree 2, 3, and 4 are depicted in Fig. 3b-d. The bias unit has also a connection to the output unit (BO connection). The total number of adjustable weights for 2,3,3-trimethylpentane is: 5 (IH connections) + 4 (BH connections) + 4 (HO connections) + 1 (BO connection) = 14.

The use of MolNet involves two phases: a learning and a prediction phase, respectively. In the learning phase the weights are adjusted with the backpropagation algorithm. The connections are modified 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. All connections from a molecule belonging to the same class are adjusted with the same quantity obtained by a summation of individual gradients and application of the usual backpropagation with momentum equation. In the prediction phase the molecular properties are computed with the weights determined in the

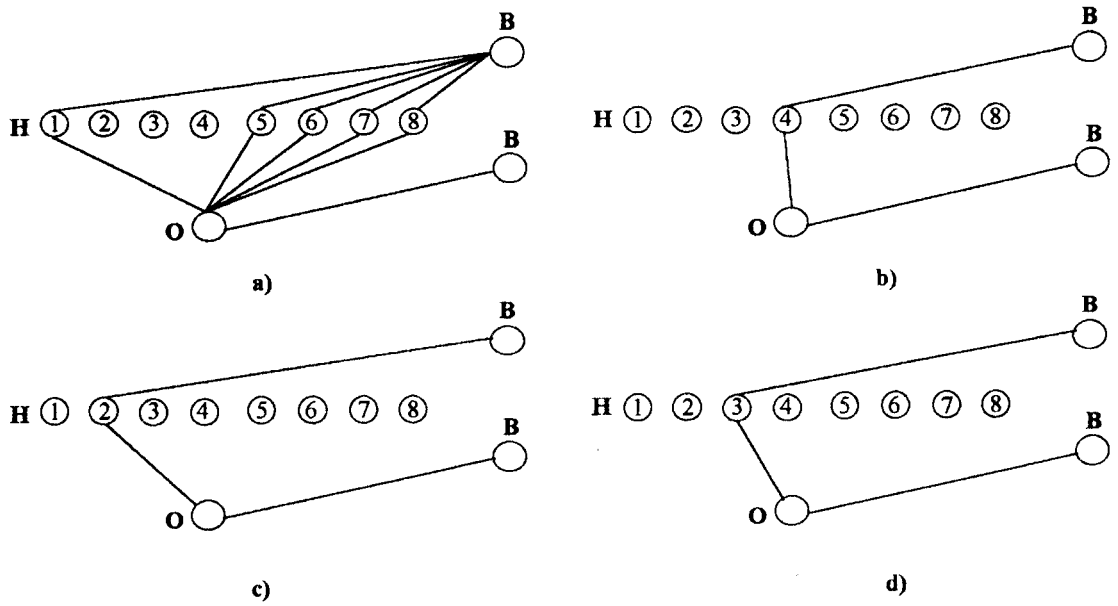


Fig. 3. – The structure of the MolNet connections between the hidden (**H**) and output (**O**) layers for 2,3,3-trimethylpentane; the bias neuron is labeled with **B**. The connections to/from atoms with the degree 1, 2, 3 and 4 are presented in a), b), c) and d), respectively.

learning phase. If the set of molecules used in the prediction phase contains bonding relationships that are absent in the molecules used in the learning phase these bonding relationships are neglected in predicting the molecular property.

The main differences between ChemNet and MolNet are presented below:

a. MolNet uses connections between the atoms with the same label in the input and hidden layers, while in ChemNet such connections are missing.

b. the bonding relationship that determines the topology of the IH connections is different in MolNet and ChemNet.

c. the structure of the HO connections in ChemNet is described only for the networks for the computation of atomic properties, while MolNet has a topology of the HO connections related to the molecular structure.

d. in ChemNet the BH connections have identical values, while in MolNet the connections from the bias to a hidden unit are classified according to the atomic number Z, the hybridization state and the degree of the atom represented by the hidden unit.

e. when computing molecular properties, ChemNet can have more than one hidden layer, while MolNet has always only one hidden layer.

MolNet operation

Data Set. MolNet was tested in a QSPR investigation of a structural determination of boiling points of alkanes. Because the scope of a QSPR study is to develop a model that gives reliable predictions for new patterns that were not used in the calibration of the model, it is necessary to estimate the prediction capabilities of the MolNet neural model, by dividing the patterns into a calibration (learning) set and a prediction (test) set. The learning set contains 109 alkanes, while the test set contains 25 alkanes between C_6 and C_{10}. The structure and experimental boiling points of the compounds used in the present investigation were taken from the literature[8] and are reported in Tables 2 and 3.

Table 2

Alkanes used in MolNet calibration, experimental boiling points, and calibration residuals for MolNet Networks using DEG and RDS input atomic descriptors

Hydrocarbon	Boiling Point °C	Residual DEG	Residual RDS	Hydrocarbon	Boiling Point °C	Residual DEG	Residual RDS
3-M-C_5	63.28	−5.19	−7.33	3,5-M_2-C_8	159.40	0.21	−0.47
2,2-M_2-C_4	49.74	−6.66	−5.28	3,6-M_2-C_8	160.80	0.80	0.32
2,3-M_2-C_4	57.99	−3.06	−3.56	4,4-M_2-C_8	157.50	1.47	−0.48
3-M-C_6	91.85	0.32	0.02	4,5-M_2-C_8	162.13	1.34	−0.14
3-E-C_5	93.48	0.20	−0.01	4-nP-C_7	157.50	0.26	−3.69
2,2-M_2-C_5	79.17	−0.33	3.42	4-iP-C_7	158.90	3.70	−0.82
2,3-M_2-C_5	89.75	1.40	2.40	2-M-3-E-C_7	161.20	3.31	1.06
2,4-M_2-C_5	80.47	−1.81	1.21	2-M-4-E-C_7	156.20	2.26	−0.91
3,3-M_2-C_5	86.04	−0.86	1.02	3-M-4-E-C_7	162.20	3.28	0.86
2,2,3-M_3-C_4	80.86	0.35	2.42	3-M-5-E-C_7	158.20	−0.88	−1.56
n-C_8	125.68	5.92	0.27	2,2,3-M_3-C_7	157.60	1.38	1.24
2-M-C_7	117.65	5.28	0.72	2,2,4-M_3-C_7	148.30	−1.35	−1.00
3-M-C_7	118.93	3.90	0.78	2,2,5-M_3-C_7	150.80	1.90	1.64
2,4-M_2-C_6	109.43	0.53	−0.09	2,2,6-M_3-C_7	148.93	3.11	3.55
2,5-M_2-C_6	109.11	4.03	1.27	2,3,3-M_3-C_7	160.20	0.54	0.35
3,3-M_2-C_6	111.97	2.23	2.02	2,3,4-M_3-C_7	159.90	−0.78	−0.47
3,4-M_2-C_6	117.73	2.56	1.78	2,3,5-M_3-C_7	160.70	2.73	2.86
3-E-2-M-C_5	115.66	2.96	1.93	2,3,6-M_3-C_7	156.00	1.64	1.17
3-E-3-M-C_5	118.27	2.16	1.88	2,4,4-M_3-C_7	151.00	−1.72	−1.63
2,2,3-M_3-C_5	109.84	1.66	1.36	2,4,5-M_3-C_7	156.50	−0.60	−0.44
2,2,4-M_3-C_5	99.24	0.58	3.86	2,4,6-M_3-C_7	147.60	−3.37	−3.59
2,3,3-M_3-C_5	114.77	2.25	1.48	2,5,5-M_3-C_7	152.80	0.55	0.74
2,3,4-M_3-C_5	113.47	2.65	3.12	3,3,5-M_3-C_7	155.68	−1.36	−0.08
2,2,3,3-M_4-C_4	106.29	0.18	−2.42	3,4,4-M_3-C_7	161.10	0.04	−0.77
2-M-C_8	143.28	6.58	−0.80	3,4,5-M_3-C_7	162.50	−1.22	−0.35
3-M-C_8	144.23	5.01	−0.01	2-M-3-iP-C_6	166.70	12.05	8.18
3-E-C_7	143.20	4.67	0.82	3,3-E_2-C_6	166.30	6.08	3.71
4-E-C_7	141.20	4.53	0.28	3,4-E_2-C_6	163.90	5.36	2.33
2,2-M_2-C_7	132.82	5.33	1.11	2,2-M_2-3-E-C_6	156.10	3.03	−1.13
2,3-M_2-C_7	140.50	4.57	0.28	2,2-M_2-4-E-C_6	147.00	−2.46	−3.93
2,4-M_2-C_7	133.20	0.97	−2.28	2,3-M_2-3-E-C_6	163.70	2.44	0.18
2,5-M_2-C_7	136.00	3.71	−0.72	2,3-M_2-4-E-C_6	160.90	1.28	−0.24
2,6-M_2-C_7	135.22	5.73	0.13	2,4-M_2-4-E-C_6	160.10	3.15	3.18
3,3-M_2-C_7	137.02	3.37	0.46	3,3-M_2-4-E-C_6	162.90	2.43	0.09
3,4-M_2-C_7	140.40	2.16	−0.53	3,4-M_2-4-E-C_6	162.10	−2.70	−3.96
3,5-M_2-C_7	135.70	−0.17	−2.61	2,2,3,3-M_4-C_6	160.31	0.11	−3.06
3-E-3-M-C_6	140.60	3.09	0.69	2,2,3,4-M_4-C_6	158.80	−0.68	−0.80
4-E-2-M-C_6	133.80	1.35	−2.03	2,2,3,5-M_4-C_6	148.40	−2.97	−3.59
2,2,4-M_3-C_6	129.91	3.95	3.03	2,2,4,5-M_4-C_6	148.88	0.00	0.37
2,2,5-M_3-C_6	124.09	3.22	0.01	2,2,5,5-M_4-C_6	137.46	−0.93	−0.99
2,3,3-M_3-C_6	137.69	2.30	−1.06	2,3,3,4-M_4-C_6	164.59	−1.25	−2.23
2,3,4-M_3-C_6	138.96	1.32	−0.33	2,3,3,5-M_4-C_6	153.10	−3.22	−3.24
2,3,5-M_3-C_6	131.36	0.98	−1.72	2,3,4,4-M_4-C_6	161.60	−1.61	−1.39
2,4,4-M_3-C_6	130.66	1.01	0.40	2,3,4,5-M_4-C_6	156.20	−4.50	−4.00
3,3,4-M_3-C_6	149.45	10.44	8.11	3,3,4,4-M_4-C_6	170.00	2.73	0.10
3,3-E_2-C_5	146.19	4.93	2.32	2,4-M_2-3-iP-C_5	157.04	1.62	−1.15
3-E-2,2-M_2-C_5	133.84	2.09	−1.57	2-M-3,3-E_2-C_5	169.70	6.58	2.95
3-E-2,3-M_2-C_5	144.70	3.88	1.04	2,2,3-M_3-3-E-C_5	169.50	5.28	0.86
2,2,3,3-M_4-C_5	140.29	2.34	−2.45	2,2,4-M_3-3-E-C_5	155.30	0.71	−2.52
2,2,3,4-M_4-C_5	133.03	0.01	−1.19	2,3,4-M_3-3-E-C_5	169.44	3.62	0.99
2,3,3,4-M_4-C_5	141.56	1.18	−1.60	2,2,3,3,4-M_5-C_5	166.05	−0.87	−4.16
3-E-C_8	166.50	3.89	2.52	2,2,3,4,4-M_5-C_5	159.29	1.09	1.01
4-E-C_8	163.64	4.00	0.82				
2,2-M_2-C_8	156.90	3.86	2.54				
2,4-M_2-C_8	155.90	−0.66	−3.09				
2,5-M_2-C_8	158.50	2.56	−0.38				
3,4-M_2-C_8	163.40	1.11	0.78				

Table 3

Alkanes used in MolNet prediction, experimental boiling points, and prediction residuals for MolNet Networks using DEG and RDS input atomic descriptors

Hydrocarbon	Boiling Point °C	Residual DEG	Residual RDS	Hydrocarbon	Boiling Point °C	Residual DEG	Residual RDS
2-M-C_5	60.27	–5.24	–6.00	2,2,4,4-M_4-C_5	122.29	3.66	6.19
n-C_7	98.40	1.94	–1.02	2,3-M_2-C_8	164.31	3.21	2.28
2-M-C_6	90.03	1.33	0.51	2,6-M_2-C_8	160.38	2.91	1.45
4-M-C_7	117.71	3.29	1.05	2,7-M_2-C_8	159.87	4.39	1.31
3-E-C_6	118.54	3.29	1.22	3,3-M_2-C_8	161.20	2.77	3.59
2,2-M_2-C_6	106.84	3.77	2.83	2-M-5-E-C_7	159.70	3.70	2.12
2,3-M_2-C_6	115.61	3.90	2.03	3-M-3-E-C_7	163.80	2.98	3.42
4-M-C_8	142.44	4.52	–0.63	4-M-3-E-C_7	163.00	2.87	1.67
4,4-M_2-C_7	134.90	2.60	–0.15	4-M-4-E-C_7	160.80	2.64	0.92
3-E-2-M-C_6	138.00	3.60	–0.63	3,3,4-M3-C_7	161.90	0.24	0.18
3-E-4-M-C_6	140.40	2.36	–0.45	2,5-M2-3-E-C_6	154.10	1.71	–1.26
2,2,3-M_3-C_6	133.58	2.00	–1.60	2,2,4,4-M4-C_6	153.80	3.31	5.53
3-E-2,4-M_2-C_5	136.73	2.92	0.30				

Number of Adjustable Parameters. Because MolNet has a variable topology, the number of adjustable parameters (connections) depends on the structure of the molecules from the learning set. In the learning set of 109 alkanes the maximum graph distance between two carbon atoms is 7, giving 8 IH connection classes. The degree of the carbon atoms is from 1 up to 4, and this gives 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 learning set and the number of adjustable weights is 6.4 and its high value indicates that there is no danger of overfitting.

Structural coding. The topological encoding of the atomic environment gives a simple and efficient description of the chemical structure in the case of highly flexible molecules. The geometry of acyclic molecules presents many local minima, making thus difficult to characterize with geometric descriptors the local and global structure. On the other hand, the identification of the ground state geometry of flexible molecules is of little use in QSPR, because at the temperature where the physical properties are determined the molecules exist in many conformations, representing local energetic minima; their identification and the description of the population of each minimum is computationally expensive, and for these reasons molecular dynamics computations are not routinely used in QSPR studies.

The input values for the units in the input layer are atomic topological descriptors namely the degree **DEG**,[19,20] the distance sum **DS**,[22,23] and the reciprocal distance sum **RDS**.[24,25] These three vertex topological indices, which were used in the present investigation, are defined in the following equations:

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

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

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

where **A** is the adjacency matrix, **D** is the distance matrix, and **RD** is the reciprocal distance matrix. Other simple atomic descriptors or vertex topological indices may be used as input data.

Learning method. The training of the ANNs was performed with the standard backpropagation method,[2] until the convergence was obtained, i.e., the correlation coefficient between experimental and calculated alkane boiling point values improved by less than 10^{-5} in 100 epochs. One epoch corresponds to one complete presentation of the molecules in the learning set, and the weights were updated after the presentation of each molecule. Random values between –0.1 and 0.1 were used as initial weights. The learning process is very sensitive to the learning rate and momentum values, and we have used small learning rates, equal to 0.05 for both the hidden and output layers. The momentum was modified between 0.30 and 0.05 for all activation functions used. Both learning rate and momentum values were maintained constant during the training phase. In all cases the networks converged in a few hundreds epochs and the results were slightly influenced by the initial random set of weights.

Activation functions. The most commonly used activation function in chemical applications of neural networks is the sigmoid that 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 was used in the present study. Both the tanh and the sigmoid activation functions are very flat when the absolute value of the argument is greater than 10 when the derivative has an extremely small value, leading to a poor sensitivity of the two activation functions to large positive or negative arguments. This behavior is 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 (e.g., the backpropagation algorithm). A linear output activation function overcomes the problems of the sigmoidal function; therefore, for the output layer it was investigated also a linear activation function. A new type of activation function is the symmetric logarithmoid,[26,27] 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.

Preprocessing of the data. Each component of the input (**DEG**, **DS**, or **RDS** vector) and output (representing the target boiling point value) patterns was 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, while for the unbounded functions (linear and symlog) the experience showed that a linear scaling improves the learning process.

Performance indicators. The performances of MolNet networks were evaluated both for the model calibration and prediction. The quality of model calibration is estimated by comparing the calculated alkane boiling points at the end of the calibration phase (BP_{cal}) with the target values (BP_{exp}), while the predictive quality was estimated by comparing the predicted (BP_{pr}) and experimental values. In order to compare the performance of different MolNet networks we have used the correlation coefficient r and the standard deviation s of the linear correlation between experimental and calculated (in calibration or prediction) boiling points: $BP_{exp} = A + B \cdot BP_{cal/pr}$.

MolNet prediction of alkane boiling points

MolNet converged always in a few hundreds epochs. Table 4 presents the calibration and prediction results obtained when MolNet was trained with **DEG** atomic descriptor as input data. The calibration correlation coefficient ranges from 0.991 to 0.994, while the prediction correlation coefficient ranges from 0.986 to 0.998. The best results are provided by a MolNet network trained with a hidden momentum $\alpha_h = 0.30$, an output momentum $\alpha_o = 0.15$ and with a linear activation function: $s_{cal} = 2.87$, $r_{cal} = 0.994$, $s_{pr} = 1.71$, $r_{pr} = 0.998$. The boiling point residuals computed with the above parameters are presented in Tables 2 and 3, column 3. Overall, the best calibration and prediction results are obtained with linear output function. MolNet networks provided with linear output function give excellent pre-

dictions for alkane boiling point with $r_{pr} > r_{cal}$, while for the networks with symlog and tanh output function $r_{pr} < r_{cal}$. This finding suggests that the linear output function is suitable for predicting alkane boiling point.

The calibration and prediction results obtained when MolNet was trained with **DS** atomic descriptor as input data are presented in Table 5. The calibration correlation coefficient ranges from 0.985 to 0.988, while the prediction correlation coefficient ranges from 0.965 to 0.976. For all MolNet networks trained with **DS** input data $r_{pr} < r_{cal}$ and the calibration and prediction results are of lower statistical quality than those obtained with **DEG** as input data. The results from Table 5 suggest that **DS** input data are not good atomic descriptors for predicting alkane boiling point with MolNet.

Table 4

MolNet calibration and prediction results for the computation of alkane boiling points 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}
1500	0.30	linear	0.30	3.49	0.991	2.84	0.995
700	0.30	linear	0.15	2.87	0.994	1.71	0.998
2000	0.30	linear	0.10	3.11	0.993	2.58	0.996
1900	0.30	linear	0.05	3.05	0.993	2.56	0.996
1800	0.15	linear	0.05	3.07	0.993	2.43	0.996
2000	0.10	linear	0.05	3.07	0.993	2.44	0.996
1900	0.05	linear	0.05	3.07	0.993	2.42	0.996
1800	0.30	symlog	0.30	3.12	0.993	4.23	0.988
1900	0.30	symlog	0.15	3.10	0.993	4.21	0.989
2000	0.30	symlog	0.10	2.83	0.994	3.65	0.991
1800	0.30	symlog	0.05	3.08	0.993	4.19	0.989
1700	0.15	symlog	0.05	3.05	0.993	4.13	0.989
1300	0.10	symlog	0.05	3.05	0.993	4.14	0.989
1400	0.05	symlog	0.05	3.04	0.993	4.13	0.989
300	0.30	tanh	0.30	3.50	0.991	4.63	0.986
400	0.30	tanh	0.15	3.50	0.991	4.64	0.986
1800	0.30	tanh	0.10	3.63	0.991	3.87	0.990
400	0.30	tanh	0.05	3.48	0.991	4.63	0.986
2000	0.15	tanh	0.05	3.17	0.993	3.73	0.991
500	0.10	tanh	0.05	3.40	0.992	4.50	0.987

Table 5

MolNet calibration and prediction results for the computation of alkane boiling points using **DS** input atomic descriptor. The notations are explained in Table 4

Epoch	Hidden momentum (α_h)	Output activation function	Output momentum (α_o)	s_{cal}	r_{cal}	s_{pr}	r_{pr}
1500	0.30	linear	0.30	4.56	0.985	7.32	0.965
700	0.30	linear	0.15	4.24	0.987	6.61	0.972
1100	0.30	linear	0.10	4.16	0.988	6.75	0.970
1300	0.30	linear	0.05	4.12	0.988	6.75	0.970
700	0.15	linear	0.05	4.07	0.988	6.22	0.975
700	0.10	linear	0.05	4.04	0.988	6.10	0.976
2000	0.05	linear	0.05	4.18	0.988	6.85	0.970
1300	0.30	symlog	0.30	4.46	0.986	6.73	0.971
2000	0.30	symlog	0.15	4.41	0.986	6.87	0.969
800	0.30	symlog	0.10	4.43	0.986	6.68	0.971
1200	0.30	symlog	0.05	4.42	0.986	6.82	0.969
1600	0.15	symlog	0.05	4.31	0.987	6.50	0.973
1800	0.10	symlog	0.05	4.27	0.987	6.46	0.973
1900	0.05	symlog	0.05	4.24	0.987	6.42	0.973
600	0.30	tanh	0.30	4.64	0.985	6.40	0.974
1800	0.30	tanh	0.15	4.53	0.985	6.75	0.971
1900	0.30	tanh	0.10	4.53	0.985	6.74	0.971
2000	0.30	tanh	0.05	4.53	0.985	6.73	0.971
1800	0.15	tanh	0.05	4.44	0.986	6.47	0.973
1900	0.10	tanh	0.05	4.42	0.986	6.41	0.973
900	0.05	tanh	0.05	4.45	0.986	6.11	0.976

Table 6

MolNet calibration and prediction results for the computation of alkane boiling points using **RDS** input atomic descriptor. The notations are explained in Table 4

Epoch	Hidden momentum (α_h)	Output activation function	Output momentum (α_o)	s_{cal}	r_{cal}	s_{pr}	r_{pr}
1700	0.30	linear	0.30	2.46	0.996	2.35	0.996
700	0.30	linear	0.15	2.49	0.996	2.60	0.996
500	0.30	linear	0.10	2.46	0.996	2.44	0.996
2000	0.30	linear	0.05	2.44	0.996	2.45	0.996
600	0.15	linear	0.05	2.41	0.996	2.37	0.996
1600	0.10	linear	0.05	2.41	0.996	2.52	0.996
700	0.05	linear	0.05	2.37	0.996	2.32	0.997
900	0.30	symlog	0.30	3.16	0.993	3.85	0.990
1100	0.30	symlog	0.15	3.16	0.993	3.84	0.991
900	0.30	symlog	0.10	3.16	0.993	3.85	0.990
1000	0.30	symlog	0.05	3.16	0.993	3.86	0.990
800	0.15	symlog	0.05	3.14	0.993	3.83	0.991
900	0.10	symlog	0.05	3.13	0.993	3.82	0.991
1000	0.05	symlog	0.05	3.13	0.993	3.81	0.991
400	0.30	tanh	0.30	3.42	0.992	4.02	0.990
500	0.30	tanh	0.15	3.43	0.992	3.97	0.990
900	0.30	tanh	0.10	3.43	0.992	3.96	0.990
500	0.30	tanh	0.05	3.41	0.992	3.99	0.990
900	0.15	tanh	0.05	3.40	0.992	3.93	0.990
600	0.10	tanh	0.05	3.38	0.992	3.92	0.990

Table 6 presents the calibration and prediction results obtained when MolNet was trained with **RDS** atomic descriptor as input data. The statistical results are close to those obtained with **DEG** input data, and much better than those obtained with **DS** atomic descriptors. The calibration correlation coefficient ranges from 0.992 to 0.996, while the prediction correlation coefficient ranges from 0.990 to 0.997. The best results are obtained with a MolNet network trained with a linear output function, and hidden and output momentum $\alpha_h = \alpha_o = 0.05$: $s_{cal} = 2.37$, $r_{cal} = 0.996$, $s_{pr} = 2.32$, $r_{pr} = 0.997$. The boiling point residuals computed with the above parameters are presented in Tables 2 and 3, column 4. The best calibration and prediction results are obtained with linear output function, while for the MolNet networks with symlog an tanh output function the results are of lower quality and $r_{pr} < r_{cal}$.

CONCLUSIONS

The MolNet neural network represents a new type of multi-layer feedforward ANN that proved to give good results in predicting alkane boiling points. MolNet changes its topology (the number of units in the input and hidden layer, and the number and type of connections) according to the molecular structure of the chemical compound presented to the network. Each non-hydrogen atom in the molecule has a corresponding unit in the input and hidden layers, while the output layer contains only one unit that provides the computed value of the molecular property.

Three atomic descriptors, namely **DEG**, **DS** and **RDS**, were used as input data, with best results obtained with the **DEG** and **RDS** indices. MolNet networks were trained with the tanh hidden activation function and three output activation functions: linear, symlog and tanh. The networks provided with linear output activation function give the best calibration and prediction results, suggesting that this output function is suitable for predicting alkane boiling points.

ACKNOWLEDGEMENT. We acknowledge the partial financial support of this research by the Ministry of Research and Technology under Grant 381 TA10 and by the Ministry of Education under Grant 5001 TB10.

REFERENCES

1. Part 8: O. Ivanciuc, *Rev. Roum. Chim.*, in press.
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. A. B. Bulsari (Ed.), "Neural Networks for Chemical Engineers", Elsevier, Amsterdam, 1995.
5. J. Devillers (Ed.), "Neural Networks in QSAR and Drug Design", Academic Press, London, 1996.
6. D. W. Elrod, G. M. Maggiora and R. G. Trenary, *J. Chem. Inf. Comput. Sci.*, **1990**, *30*, 477–484.
7. D. W. Elrod, G. M. Maggiora and R.G. Trenary, *Tetrahedron Comput. Methodol.*, **1990**, *3*, 163–174.
8. A. A. Gakh, E. G. Gakh, B. G. Sumpter and D. W. Noid, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 832–839.
9. A. T. Balaban, S. C. Basak, T. Colburn and G. D. Grunwald, *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 1118–1121.
10. D. Cherqaoui and D. Villemin, *J. Chem. Soc. Faraday Trans.*, **1994**, *90*, 97–102.
11. D. Cherqaoui, D. Villemin, A. Mesbah, J.-M. Cense and V. Kvasnička, *J. Chem. Soc. Faraday Trans.*, **1994**, *90*, 2015– 2019.
12. F. R. Burden, *Quant. Struct.-Act. Relat.*, **1996**, *15*, 7–11.
13. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye and J. P. Doucet, *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 644–653.
14. O. Ivanciuc, J.-P. Rabine, D. Cabrol-Bass, A. Panaye and J. P. Doucet, *J. Chem. Inf. Comput Sci.*, **1997**, *37*, 587–598.
15. O. Ivanciuc, J.-P. Rabine and D. Cabrol-Bass, *Comput. Chem.*, **1997**, *21*, 437–443.
16. D. B. Kireev, *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 175–180.
17. I. I. Baskin, V. A. Palyulin and N. S. Zefirov, *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 715–721.
18. O. Ivanciuc, MolNet Neural Network Application in Structure-Property Studies, The 23rd Chemistry Conference, 8-10 October 1997, Climneti, Vlcea, Romania.
19. M. V. Diudea, and O. Ivanciuc, "Molecular Topology", Comprex, Cluj, Romania, 1995.
20. O. Ivanciuc, and A.T. Balaban, "Graph Theory in Chemistry", in: "Encyclopedia of Computational Chemistry", Ed.: P. v. R. Schleyer, Wiley, 1998.
21. B. Mohar and T. Pisanski, *J. Math. Chem.*, **1988**, *2*, 267–277.
22. A.T. Balaban, *Chem. Phys. Lett.*, **1982**, *89*, 399–404.
23. A.T. Balaban, *Pure Appl. Chem.*, **1983**, *55*, 199–206.
24. O. Ivanciuc, *Rev. Roum. Chim.*, **1989**, *34*, 1361–1368.
25. O. Ivanciuc, T.-S. Balaban and A. T. Balaban, *J. Math. Chem.*, **1993**, *12*, 309–318.
26. A. B. Bulsari and H. Saxén, *Neurocomputing*, **1991**, *3*, 125–133.
27. A.B. Bulsari and H. Saxén, *Neural Network World*, **1991**, *4*, 221–224.