Chapter 8

NEW NEURAL NETWORKS FOR STRUCTURE-PROPERTY MODELS

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

1. INTRODUCTION

The growing interest in the application of Artificial Neural Networks (ANN)[1-4] in chemistry,[5-8] in chemical engineering,[9] and in biochemistry[10] is a result of their demonstrated superiority over the traditional models. Neural networks were used in quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR) studies to predict various physical, chemical, and biological properties of organic compounds. Conventional QSPR and QSAR require the user to specify the mathematical function of the model. If these functions are highly nonlinear then considerable mathematical and numerical expertise is needed to obtain significant models. In recent years this problem was solved with ANN that generate highly non-linear models between input and output variables; an important advantage of ANN is the fact that the mathematical form of the relationship between the input and output variables is not specified by the user. An ANN may be represented as an assembly of units (artificial neurons) that performs some basic mathematic operations and that receive and send signals through connections. Each connection between two neurons has an adjustable weight that modifies the signal send through that connection. The artificial neuron is a simplified model of a neuron having four basic functions: receives input signals from the environment or from the neurons in the previous layer; makes a

summation of the inputs signals; the result of summation is passed to the activation function and transformed with a mathematical function; the result of the activation function is used by a transfer function to produce an output value that is send to the neurons in the next layer. The most important activation functions are: linear, $Act(x) = x$; sigmoid, $Act(x) = 1/(1 + e^{-x})$; hyperbolic tangent, $Act(x) = tanh(x)$; symmetric logarithmoid, $Act(x) = \text{sign}(xx) \ln(1+|x|)$; bell, $Act(x) = 1/(1 + x^2)$.[11-15]

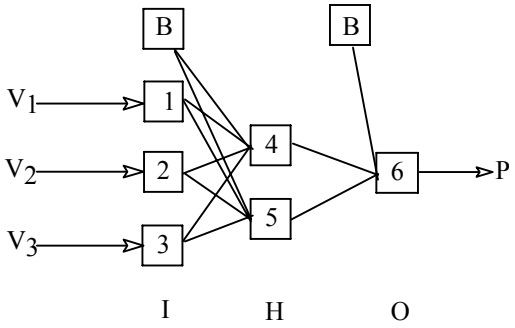


Figure 1. The topology of a multilayer feed-forward neural network with three layers

In multi-layer feedforward (MLF) networks, usually used in QSPR and QSAR studies, the neurons are arranged in distinct layers, with connections between the neurons in different layers; the structure of a simple MLF network with three layers is presented in Figure 1. The input layer I receives the input signals from the environment and sends them to the next layer. The middle layer H is called the hidden layer because it is not connected directly to the environment, but receives signals from the input layer, transforms them, and sends information to the output layer or to the next hidden layer. The output layer O receives signals from the hidden layer, processes them, and sends to the environment the result computed by the network. Each neuron in the hidden and output layers is connected to a bias neuron B. An MLF contains a number of input neurons equal to the number of descriptors that describe the structure of each chemical compound, and a number of output neurons equal to the number of physical, chemical, or biological properties modeled by the ANN. The network in Figure 1 has three input neurons, meaning that the chemical structure is represented with three descriptors, denoted here V_1, V_2, and V_3. When a pattern representing in numerical form the structure a certain molecule is presented to the network, descriptor V_1 goes to input neuron 1, V_2 goes to input neuron 2, and V_3 goes to input neuron 3. In this way the information regarding the chemical structure of a molecule is entered into the ANN. The hidden layer is formed by neurons 4 and 5, and the output layer contains neuron 6. The output neuron 6 offers the computed value of the property P for the compound whose structure was encoded in descriptors V_1, V_2, and V_3. We have to note here that the multilinear regression model, widely used in QSPR and QSAR, is equivalent with a multi-layer feedforward network without hidden layer and with one output neuron provided with linear activation function.

An important problem for the QSPR and QSAR applications of ANN remains the numerical representation of the chemical structure used as input by the ANN; a recent review presents the molecular graph descriptors used as input data for ANN models.[16] The molecular graph descriptors[17-21] represent an important class of structural descriptors for QSPR and QSAR studies; these descriptors present important advantages when compared with other classes of descriptors because they can be easily computed from the molecular graph and they do not require the computation of the molecular geometry. For example, simple graph descriptors (the number of carbon atoms and the number of pairs of methyl groups separated by 3 up to 8 carbon-carbon single bonds) were used to obtain an ANN model for six physical properties of alkanes.[22]

The usual structure of an MLF network is not dependent on the chemical structure of the molecules from the calibration and prediction sets. Three new neural networks were defined in order to encode into their topology the chemical structure of each compound presented to the network: the Baskin-Palyulin-Zefirov neural device,[23] ChemNet defined by Kireev,[24] and MolNet introduced by Ivanciuc.[25,26] All the above neural networks change their 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 represented as a molecular graph that is used to set the ANN topology. In this review we present the rules that define the three neural models together with examples of network generation from the molecular graph.

2. THE BASKIN, PALYULIN, ZEFIROV NEURAL DEVICE

The Baskin, Palyulin, Zefirov (BPZ)[23] neural device is a special MLF neural network that generates nonlinear QSPR and QSAR correlations using as input data local invariants describing the atomic environment or bonding relationships. The structure of the BPZ neural device, presented in Figure 2, is constructed by analogy with a biological vision system and contains a sensor field, a set of eyes, and a brain. A chemical structure modeled by the BPZ neural device is represented as a molecular matrix that is superimposed over a sensor field. The sensor field receives the network input from the numerical representation of the chemical structure. The information received by the sensor field is transmitted to a set of eyes that transform it in several MLF networks, and sends signals to the next block, the brain. All output signals from the eyes are processes in the brain with the aid of an MLF ANN that offers at its output the computed values of the investigated properties. According to a set of rules the structure of the sensor field and of the eyes depends on the chemical nature of the atoms and their bonding relationships for each molecule presented to the neural device.

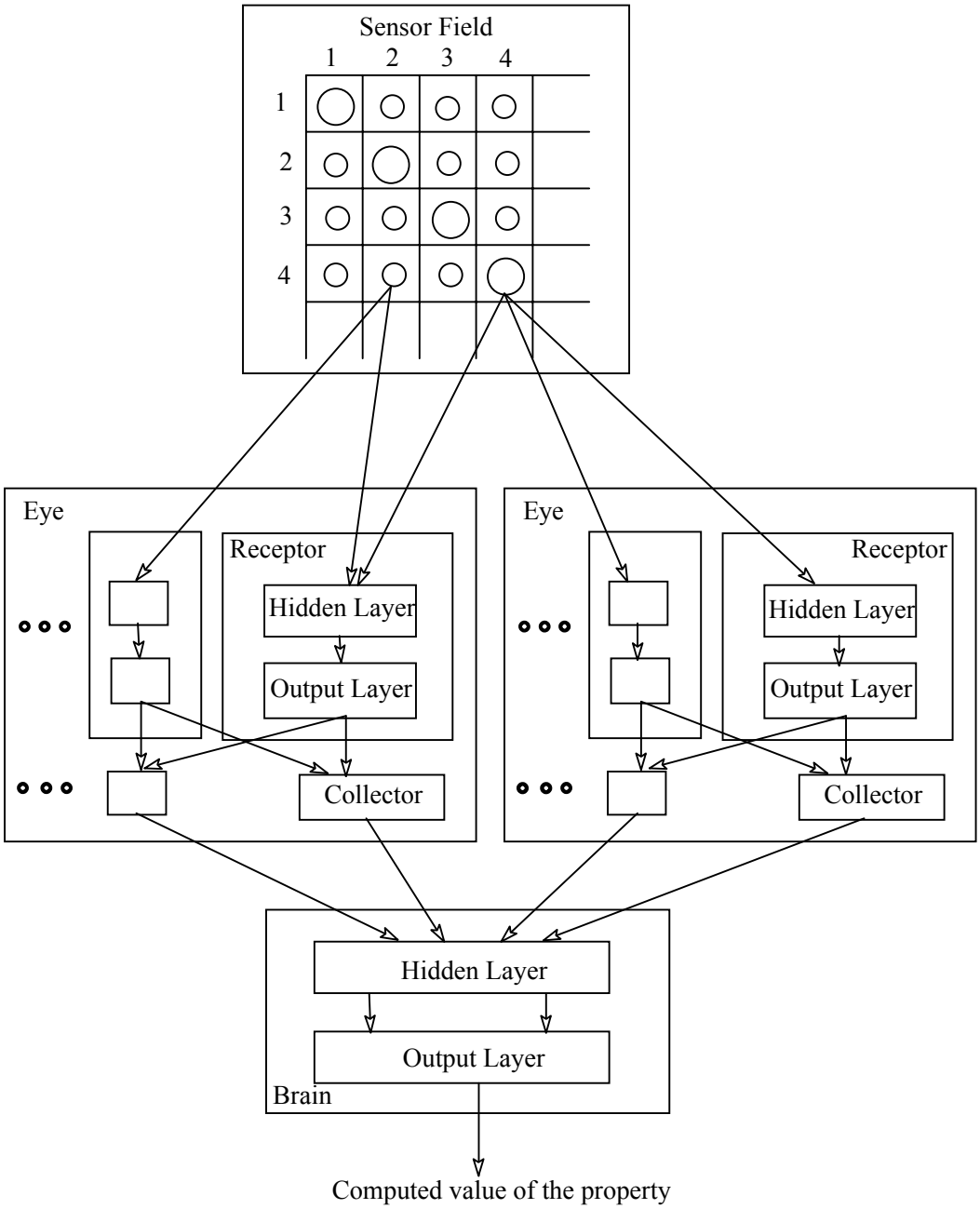


Figure 2. The structure of the Baskin, Palyulin, Zefirov neural device.

All input neurons in the sensor field receive specific structural elements from a chemical compound and send this information to the neural networks situated in the eyes. The neurons that form the eyes and the brain receive signals, sum them, transform the sum, and send output signals. These neurons are connected to a bias neuron; in the mathematical equations that describe their function the connection from the bias neuron

is represented as a threshold. The three building blocks that form the BPZ neural device are presented in detail below.

Using a matrix representation, the chemical structure of each organic compound presented to the BPZ neural device is encoded into the sensor field. The atoms are labeled from 1 to N, where N is the number of non-hydrogen atoms from the current molecule. The molecular structure is represented by a $N \times N$ matrix in which the ii-th element (situated on i-th position of the diagonal) corresponds to a certain property of the ith atom in the current molecule, and the ij-th element (situated at the intersection of the ith row with the j-th column) is a numerical value representing some relationship between atoms i and j from the current molecule. This molecular matrix is symmetric, i.e. the ij-th element is identical with the ji-th element. The atomic properties that can be used on the main diagonal of this matrix are the principal quantum number, the electronegativity, the atomic charge, the number of valence electrons, the number of hydrogen atoms attached to the atom. The nondiagonal matrix elements represent various interatomic relationships, such as connectivity information, the interatomic distance or the bond order. The sensor field is formed by input neurons arranged in a square $N \times N$ matrix. Each neuron corresponds to an element in the molecular matrix from the current molecule. A sensor neuron receive the value of the corresponding molecular matrix element and send the structural information through weighted connections to the eyes.

The BPZ neural device contains two or more eyes, each one receiving specific information about the chemical structure of the current molecule. An eye contains a set of identical receptors and one or more collectors. All receptors situated in an eye are identical, containing the same number of neurons, the same weights for connections between identical pairs of neurons, and equal thresholds for identical neurons situated in different receptors. A receptor is an MLF network that receives signals from the sensor field, processes them in a hidden layer, and from the output layer sends signals to the collector(s). The hidden layer can have any number of neurons, but the output layer always contains one neuron that sends signals to the collectors from the same eye. Each eye receptor is connected to a specific field from the sensor field that contains a part of the information related to the chemical structure. For each receptor there is an ordered vector $\{v_1, v_2, ..., v_i, \ldots v_n\}$, where n is the number of atoms sending signals to the receptor and v_i is the label of the corresponding atom in a molecule. This specific vector is called a receptor identifier and represents the molecular substructure perceived by a specific receptor. The number of receptors situated in an eye is equal to the number of different receptor identifiers that can be generated, i.e. $N!/(N-n)!$, where N is the number of non-hydrogen atoms in the molecule encoded in the sensor field, and n is the number of atoms in a receptive field. A receptor identifier containing n atoms is called n-atomic. A molecule that contains three non-hydrogen atoms can be analyzed by the BPZ neural device in three ways: with three receptors receiving signals from the one-atomic receptor identifiers {1}, {2}, and {3}; with six receptors receiving signals from the two-atomic receptor identifiers {1, 2}, {1, 3}, {2, 1}, {2, 3}, {3, 1}, and {3 ,2}; with six receptors receiving signals from the three-atomic receptor identifiers {1, 2, 3}, {1, 3, 2}, {2, 1, 3}, {2, 3, 1}, {3, 1, 2}, and {3, 2, 1}.

The signals processed in each receptor from an eye are send to one or more collectors from the same eye. Each collector contains a neuron that receives signals, sums them, transforms the sum, and sends a signal to the brain of the BPZ neural device. The brain is a feed-forward neural network that receives signals from every collector, processes them in a hidden layer, and from the output layer sends signals that represent the computed value of the investigated molecular property. The hidden layer can contain any number of neurons, but the output layer always contains a number of neurons equal to the number of molecular properties modeled with the BPZ neural device. Each output neuron corresponds to one and only one molecular property.

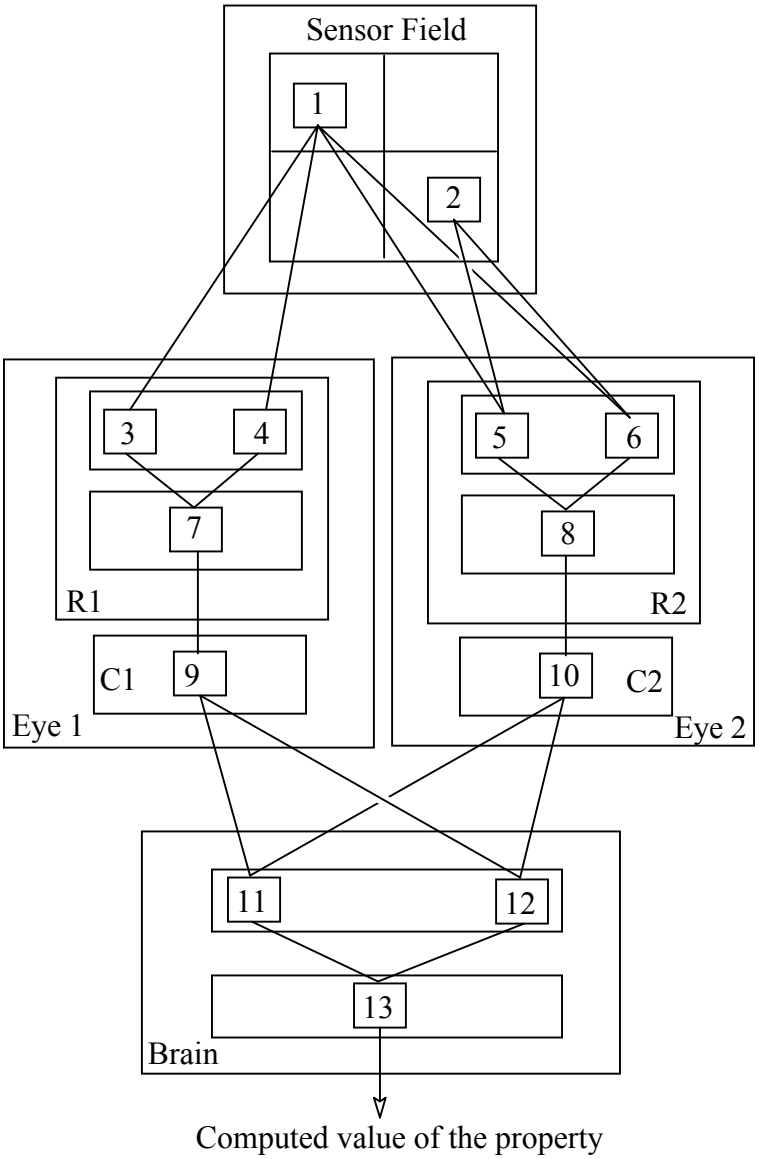


Figure 3. A minimal configuration of the Baskin, Palyulin, Zefirov neural device.

The BPZ neural device that contains all receptors needed for the perception of a molecule represents the configuration for that molecule. A BPZ neural device with a minimal configuration, presented in Figure 3, has one receptor in each eye and contains only mutually independent adjustable parameters. A minimal configuration can be used as a template for the construction of the BPZ neural device for any molecule presented to the neural device and encoded in the sensor field. In this cqse, the sensor field is a square matrix containing only atomic sensors situated on the diagonal of the matrix. The minimal configuration has two eyes with different functions: eye 1 receive signals from individual atoms, while eye 2 receives from the sensor field signals corresponding to pairs of bonded atoms. The signals from the sensor neuron 1 go to neurons 3 and 4 that represent the hidden layer of the receptor R1 from Eye 1. The processed signals are send to the output neuron 7. The collector C1 has one neuron, 9, that gets the signal from the eye receptor and sends it to the brain. From the neurons in the sensor field Eye 2 receives signals from pairs of bonded atoms: neurons 1 and 2 send two signals to receptor neuron 5, and another pair of signals to receptor neuron 6. The signal flow in Eye 2 is similar with that from Eye 1: from the input neurons 5 and 6 to the output neuron 8 and then to the collector neuron 10. The brain has an input layer that contains neurons 11 and 12, and an output layer containing the neuron 13. The brain input neurons 11 and 12 receive signals from the two eyes collector neurons 9 and 10, respectively. Brain output neuron 13 offers the computed molecular property of the molecule encoded in the sensor field.

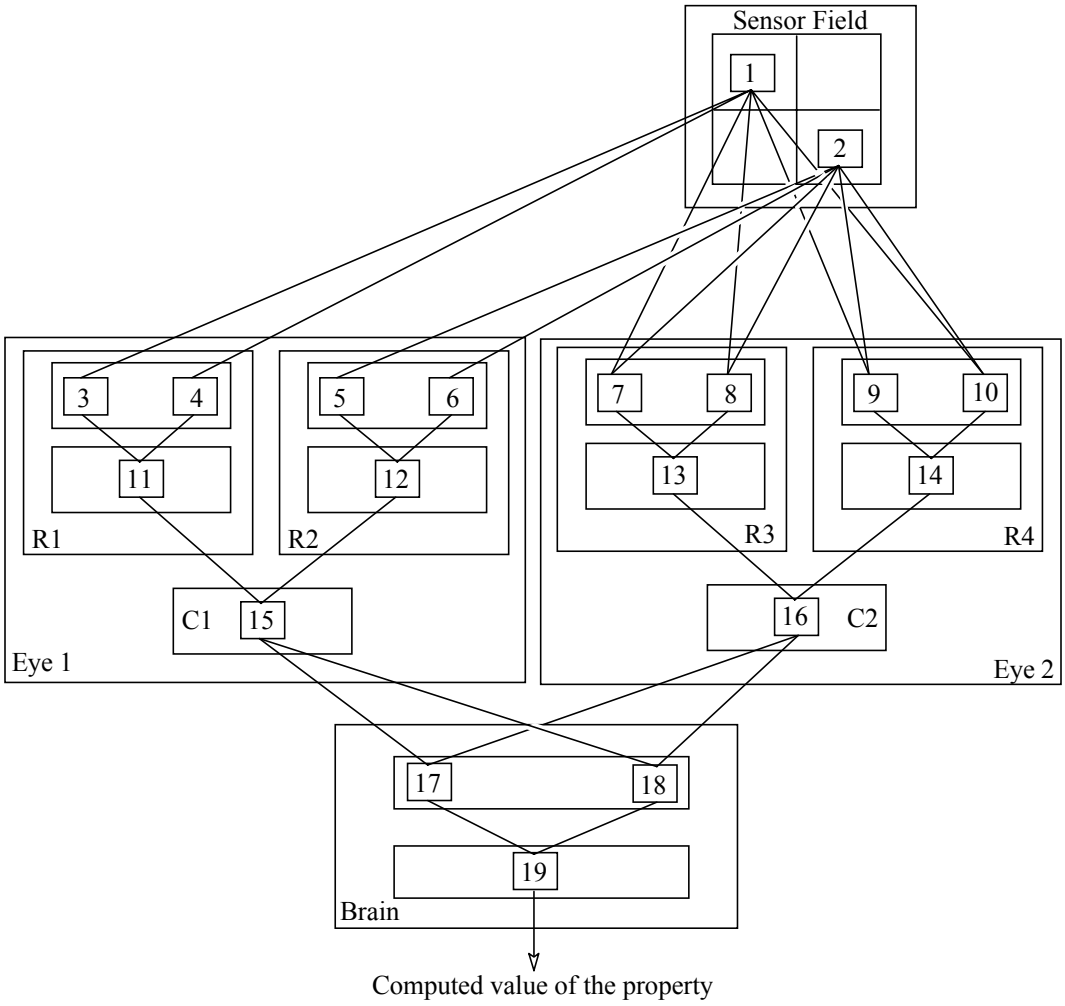


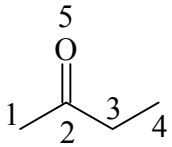
Figure 4. The structure of the Baskin, Palyulin, Zefirov neural device for ethane.

The structure of the BPZ neural device for ethane is presented in Figure 4. Ethane is considered here as its hydrogen-depleted molecular graph. The threshold of a neuron i is denoted with θ_i and the weight of the connection between neurons i and j is denoted with w_{ij}. According to the rules for generating a BPZ neural device, certain connections weights and neuron thresholds have identical values. In this way, the number of adjustable parameters is much lower than the number of connections between neurons. For ethane, the receptor field is a square 2×2 matrix, containing only atomic sensors (neurons 1 and 2) situated on the diagonal of the matrix, corresponding to the two carbon atoms in ethane. The input value for the sensor neurons can be any property of the corresponding atom; if this property is the atomic number Z, both sensor neurons in ethane receive the value 6. The neural device presented in Figure 3 consists of two eyes, Eye 1 and Eye 2, with different functions: each receptor situated in Eye 1 receives signals from only one sensor neuron representing an atom, while each receptor from Eye 2 receives signals from two-atomic receptive fields corresponding to atoms that have a

bond between them in the molecule encoded in the sensor field. Each eye has two receptors and one collector. The ethane BPZ neural device is generated from the minimal configuration neural device presented in Figure 3, and as one can see from Figure 4, each receptor has two hidden and one output neuron. From the BPZ neural device generation algorithm it results that the following eye neurons thresholds are identical: $\theta_3 = \theta_5$, $\theta_4 = \theta_6$, $\theta_7 = \theta_9$, $\theta_8 = \theta_{10}$, $\theta_{11} = \theta_{12}$, and $\theta_{13} = \theta_{14}$. Some connections between neurons in the sensor field and neurons in the receptors input layers have identical weights: $w_{1,3} = w_{2,5}$, $w_{1,4} = w_{2,6}$, $w_{1,7} = w_{2,9}$, $w_{1,9} = w_{2,7}$, $w_{1,8} = w_{2,10}$, and $w_{1,10} = w_{2,8}$. One finds also identical weights for certain connections between neurons in the receptor input and output layers: $w_{3,11} = w_{5,12}$, $w_{4,11} = w_{6,12}$, $w_{7,13} = w_{9,14}$, and $w_{8,13} = w_{10,14}$. Finally, the following connections between receptors and collectors have identical weights: $w_{11,15} = w_{12,15}$, and $w_{13,16} = w_{14,16}$. Each eye contains a single collector containing one neuron, namely neurons 15 and 16, and each eye sends signals to the brain. The BPZ neural device was successfully used in the prediction of alkanes boiling points, hydrocarbons viscosity, heat of evaporation, and density, cyclohexane heat of solvation, organic compounds polarizability, and gases anesthetic pressure.

3. CHEMNET

ChemNet is an MLF neural network that encodes in its topology the structure of each molecule presented to the network, and computes an atomic property, such as the atomic charge.[24] Each molecule is represented by its labeled hydrogen-depleted molecular graph. For each atom *i* from the molecular graph ChemNet has two corresponding neurons with the same label *i*, one in the input layer and the second one in the hidden layer. The output layer has only one neuron, representing the atom from the molecular graph whose property is computed with ChemNet. The network contains also a bias neuron, connected to the hidden and output neurons. ChemNet changes the number and significance of the neurons with each molecule presented to the network. The connections between the input and hidden layers correspond to the bonding relationships between pairs of atoms, i.e. two pairs of atoms separated by the same number of bonds have identical connection weights.



1

As an example of ChemNet structure we present the network generation for butanone **1**. In ChemNet the bonding relationship that determines the type of the connections between input and hidden layers considers only the number of bonds separating a pair of atoms. These relationships can be determined from the distance matrix of the molecular

graph computed by considering that all atoms are carbons and all bonds are single. The distance matrix of the molecular graph **1**, **D**(**1**), is:

D(1)

	1	2	3	4	5
1	0	1	2	3	2
2	1	0	1	2	1
3	2	1	0	1	2
4	3	2	1	0	3
5	2	1	2	3	0

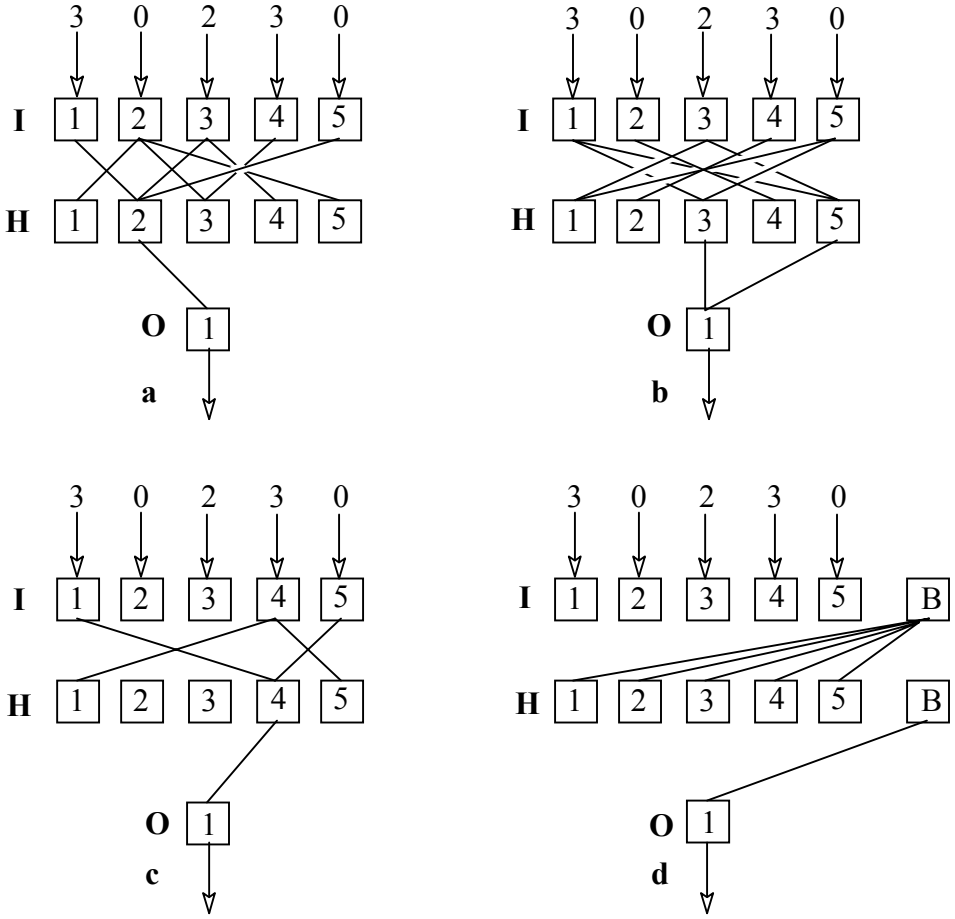


Figure 5. The ChemNet topology for butanone **1** that computes a property of atom 1. Each neuron in the input (**I**) and hidden (**H**) layers corresponds to the atom with the same label from the molecular graph **1**. The connections between atoms situated at distances 1, 2, and 3 are presented in **a**, **b**, and **c**, respectively, while the connections from the bias neuron B are presented in **d**. Input values represent the number of hydrogen atoms attached to each atom in the molecular graph **1**.

The molecular graph of butanone has 5 atoms and the corresponding ChemNet has 5 input and 5 hidden neurons. Each neuron from the input and hidden layers of ChemNet corresponds to the atom with the same label in the molecular graph **1**, as presented in Figure 5a-d; for example, the oxygen atom is represented by the input and hidden neurons labeled with 5. All pairs of atoms in **1** that are separated by the same number of bonds are characterized by input-hidden connections with identical weights. A bonding relationship between two atoms i and j corresponds to two identical input-hidden connections: the first from input neuron i to hidden neuron j, and the second from input neuron j to hidden neuron i. In **1** there are 4 pairs of atoms separated by one bond, as can be seen also from the distance matrix **D**(**1**). In Figure 5a we present the ChemNet connections corresponding to pairs of atoms separated by one bond, giving the following equalities between the corresponding weights: $w_{1,2} = w_{2,1} = w_{2,3} = w_{3,2} = w_{3,4} = w_{4,3} = w_{2,5} = w_{5,2}$. The ChemNet connections corresponding to pairs of atoms separated by two bonds are presented in Figure 5b. From the molecular graph of **1** and its corresponding distance matrix it follows that $w_{1,3} = w_{3,1} = w_{1,5} = w_{5,1} = w_{2,4} = w_{4,2} = w_{3,5} = w_{5,3}$. The 4 input-hidden connections corresponding to pairs of atoms separated by three bonds are presented in Figure 5c; they give the following equalities between the corresponding weights: $w_{1,4} = w_{4,1} = w_{4,5} = w_{5,4}$. All connections from the bias neuron to the 5 hidden neurons, presented in Figure 5d, have the same weight: $w_{B,1} = w_{B,2} = w_{B,3} = w_{B,4} = w_{B,5}$.

In Figure 5 we present the ChemNet structure for butanone when atom 1 is the output atom, i.e. the network computes an atomic property of the atom 1. In this case, the output layer contains one neuron representing the atom 1 from butanone. The type of a hidden-output connection joining a hidden neuron i with an output neuron j, denoted v_{ij}, is determined by the number of bonds between atoms i and j. There is one hidden-output connection corresponding to atoms one bond away from atom 1, representing the bond between atoms 2 and 1, as presented in Figure 5a. The ChemNet hidden-output connections corresponding to atoms two bonds away from atom 1 are presented in Figure 5b. From the molecular graph of **1** and its corresponding distance matrix it follows that two hidden-output connections have identical weights, namely $v_{3,1} = v_{5,1}$. The connection between atoms 4 and 1, separated by three bonds, is depicted in Figure 5c.

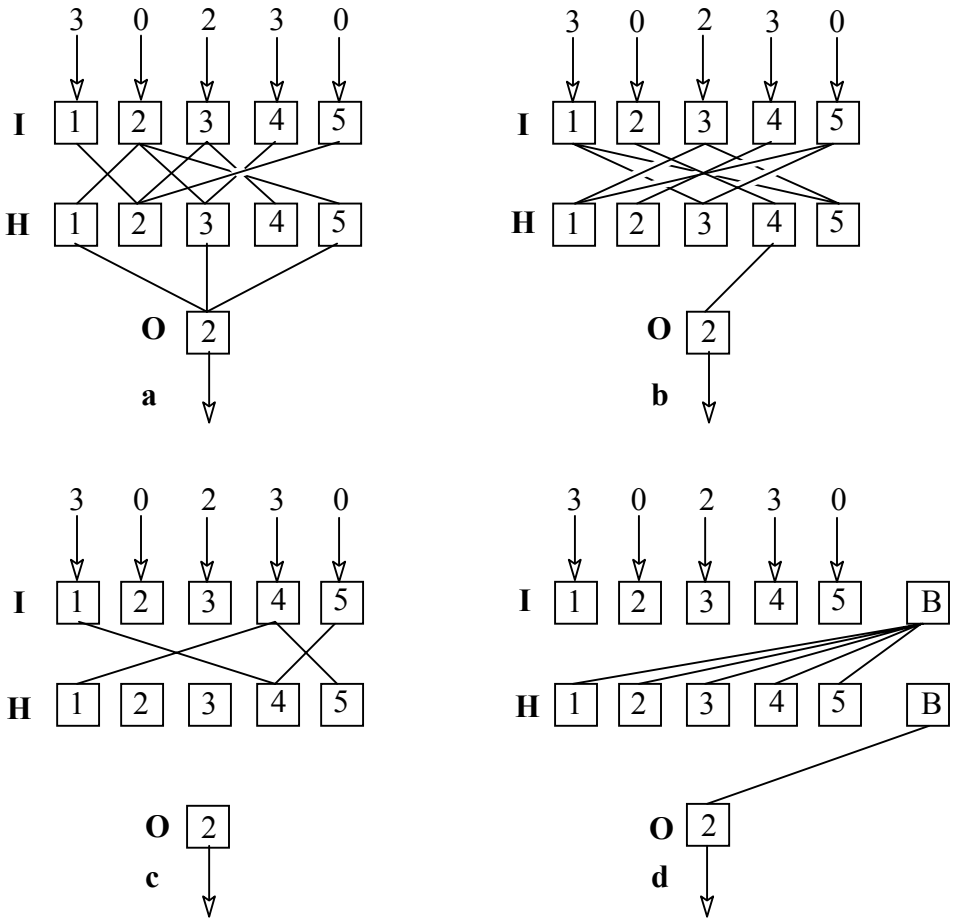


Figure 6. The ChemNet topology for butanone **1** that computes a property of atom 2. Each neuron in the input (**I**) and hidden (**H**) layers corresponds to the atom with the same label from the molecular graph **1**. The connections between atoms situated at distances 1, 2, and 3 are presented in **a**, **b**, and **c**, respectively, while the connections from the bias neuron B are presented in **d**. Input values represent the number of hydrogen atoms attached to each atom in the molecular graph **1**.

The ChemNet topology that computes an atomic property for atom 2 in butanone is given in Figure 6. The input and hidden layers, together with the connections between them, are identical with those from the ChemNet network generated for the output atom 1, as can be seen by comparing Figures 5 and 6. Therefore, we will explain in detail only the connections between the hidden and output layers. The ChemNet hidden-output connections corresponding to atoms one bond away from atom 2 are presented in Figure 6a, and correspond to three hidden-output connections with identical weights: $v_{1,2} = v_{3,2} = v_{5,2}$. There is one hidden-output connection corresponding to atoms two bonds away from atom 2, representing the path between atoms 4 and 2, as presented in Figure 6b, and there is no atom three bonds away from atom 2, as can be seen from Figure 6c. The connections from the bias neuron to the 5 hidden neurons, and to the output neuron

representing atom 2 are depicted in Figure 6d. ChemNet was applied for the computation of AM1 atomic charges.

4. MOLNET

MolNet is a multi-layer feedforward neural network that can be used to compute molecular properties on the basis of chemical structure descriptors.[25,26] A specific feature of MolNet is that the number of neurons and the connections between them depend on the chemical structure of each molecule presented to the network. Molecules are represented by the corresponding hydrogen-suppressed molecular graph. Each non-hydrogen atom in a molecule has a corresponding neuron in the input and hidden layers. The number of neurons in the input and hidden layers is equal to the number of atoms from the molecular graph. The output layer has only one neuron, providing the calculated value of the molecular property under investigation. The network has a bias neuron, connected to all hidden and output neurons.

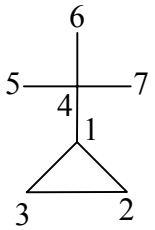
The connection between an input and a hidden neuron characterizes the bonding relationship between the corresponding atoms. Input-hidden connections corresponding to the same bonding relationship between two atoms, either in the same molecule or in different molecules, belong to the same connection class and they have identical weights. In MolNet the bonding relationship considers the shortest path between a pair of atoms. Two identical paths contain the same ordered sequence of atoms and bonds; such paths belong to the same connection class. A neuron that corresponds to an atom i in the input layer is connected to the neuron 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.

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 neurons. We have to point here that even for neurons 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 neurons in the same class. Each neuron 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 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, N connections from the bias neuron to the hidden neurons, and one connection from the bias neuron to the output neuron. Some connections may have identical weights according to the partitioning schemes described above. This implies that for MolNet the number of adjustable parameters is much smaller than the number of connections. When

a molecule is presented to MolNet input neuron *i* receives a signal representing an atomic property computed for the atom *i* of the respective molecular graph. Any vertex invariant computed from the structure of the molecular graph can be used as input for MolNet.

As an example of MolNet generation we consider *t*-butylcyclopropane **2** whose molecular graph is presented below:



2

In the molecular graph of an alkane, the topological distance between two vertices *i* and *j*, d_{ij}, is equal to the number of edges (corresponding to carbon-carbon single bonds) on the shortest path between the vertices *i* and *j*. Distances d_{ij} are elements of the distance matrix of a molecular graph *G*, **D** = **D**(*G*). The distance matrix of the molecular graph of **2**, **D**(**2**), is:

D(**2**)

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

Molecule **2** contains 7 carbon atoms. Each carbon atom from the molecular graph **2** has a corresponding neuron with the same label in the input and hidden layers of MolNet, as presented in Figure 7a-d. The distance matrix of *t*-butylcyclopropane has 4 classes of topological distances: 7 distances 0; 7 distances 1; 8 distances 2; 6 distances 3. The 4 types of graph distances correspond to four IH connection types or parameters that are adjusted during the calibration phase. A bonding relationship between two atoms *i* and *j* corresponds to 2 identical IH connections: the first from input neuron *i* to hidden neuron *j*, and the second from input neuron *j* to hidden neuron *i*. As an example of identical IH connections, consider the 6 pairs of atoms: 2 and 5; 2 and 6; 2 and 7; 3 and 5; 3 and 6; 3 and 7. The graph distance between the atoms in the above pairs is 3. Therefore, for

molecule **2** there are 12 IH connections with identical weights between the above 6 pairs of atoms, as depicted in Figure 7d. These 12 connections have an identical weight and correspond to the parameter for two carbon atoms situated at distance 3.

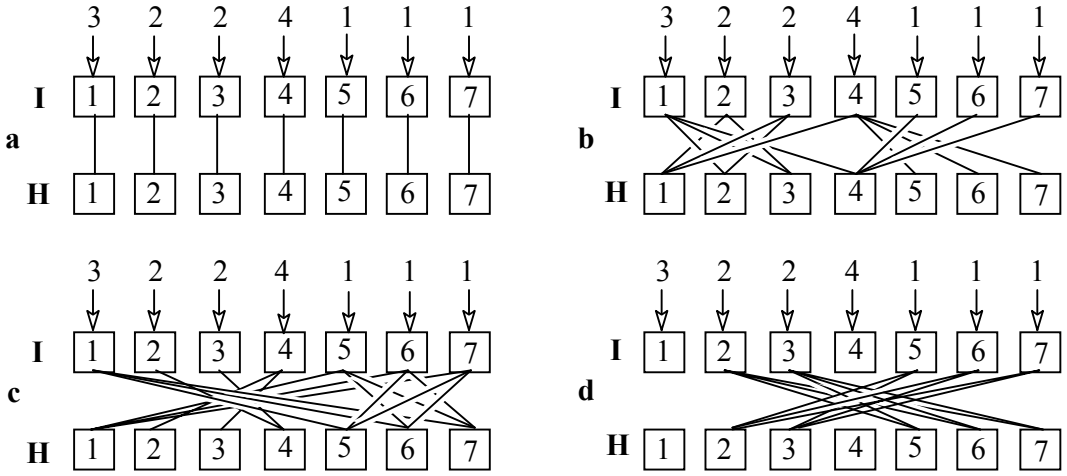


Figure 7. The structure of the MolNet IH connections between the input (**I**) and hidden (**H**) layers for *t*-butylcyclopropane **2**; each neuron corresponds to the carbon atom with the same label from the molecular graph **2**. The connections between atoms with the same label are presented in **a**, while the connections between atoms situated at distances 1, 2, and 3 are presented in **b**, **c**, and **d**, respectively. Input values represent vertex degrees.

In Figure 7a-d we present the structure of IH connections according to the classes of identical weights: there are 7 connections corresponding to distance 0 (Figure 7a), which in our case have identical weights because all non-hydrogen atoms are carbon atoms $w_{1,1} = w_{2,2} = w_{3,3} = w_{4,4} = w_{5,5} = w_{6,6} = w_{7,7}$; the 7 pairs of atoms situated at distance 1 correspond to the 14 connections with identical weights in Figure 7b: $w_{1,2} = w_{2,1} = w_{1,3} = w_{3,1} = w_{1,4} = w_{4,1} = w_{2,3} = w_{3,2} = w_{4,5} = w_{5,4} = w_{4,6} = w_{6,4} = w_{4,7} = w_{7,4}$; Figure 7c presents the 16 connections between the 8 pairs of atoms situated at distance 2, with the following equalities between connection weights: $w_{1,5} = w_{5,1} = w_{1,6} = w_{6,1} = w_{1,7} = w_{7,1} = w_{2,4} = w_{4,2} = w_{3,4} = w_{4,3} = w_{5,6} = w_{6,5} = w_{5,7} = w_{7,5} = w_{6,7} = w_{7,6}$; the 12 connections corresponding to the 6 pairs of atoms situated at distance 3 are presented in Figure 7d: $w_{2,5} = w_{5,2} = w_{2,6} = w_{6,2} = w_{2,7} = w_{7,2} = w_{3,5} = w_{5,3} = w_{3,6} = w_{6,3} = w_{3,7} = w_{7,3}$.

The connections between the hidden and output layers (the HO connections) are determined by the structure of the molecule presented to MolNet. In alkanes the HO connections are classified according to the degree of the carbon atoms: hidden neurons representing atoms with identical degrees are linked to the output neuron with connections having identical weights. Because the molecular graph of *t*-butylcyclopropane contains three atoms with degree 1, two atoms with degree 2, one atom with degree 3, and one atom with degree 4, the 7 HO connections belong to four classes (i.e. adjustable parameters). The connections between the bias neuron and the neurons 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 **2** four adjustable

weights. For example, there are 3 atoms with degrees 1 in *t*-butylcyclopropane, namely 5, 6, and 7. Figure 8a presents their connections from the bias neuron and to the output neuron. From the rule used to generate the MolNet connections it follows that the 3 atoms with degrees 1 have HO connections with identical weights. If we denote with $v_{i,o}$ the weight of the connection between hidden neuron i and output neuron o, and with $v_{b,i}$ the weight of the connection between the bias neuron b and the hidden neuron i, we have the following equalities: $v_{5,o} = v_{6,o} = v_{7,o}$ and $v_{b,5} = v_{b,6} = v_{b,7}$. Because the 3 atoms are also topologically equivalent, the signal send to the output neuron is identical. The same Figure 8a depicts the 3 identical BH connections to the 3 atoms with degrees 1. As a consequence, the signal received from the bias by the neurons 5, 6, and 7 is identical.

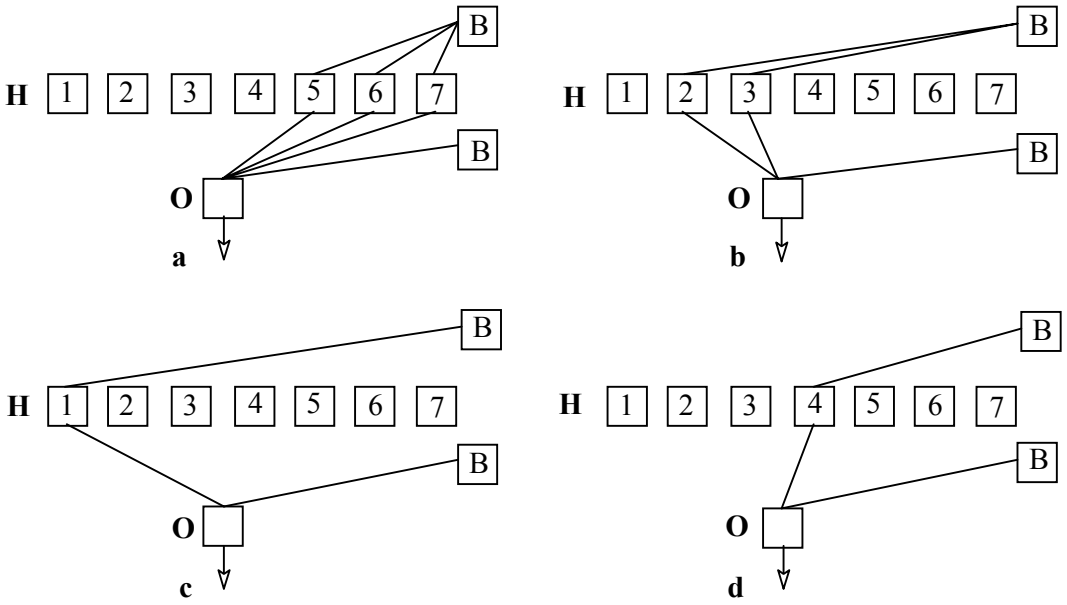


Figure 8. The structure of the MolNet connections between the hidden (**H**) and output (**O**) layers for *t*-butylcyclopropane **2**; the bias neuron is labeled with **B**. The BH and HO connections to/from hidden neurons representing atoms with the degree 1, 2, 3, and 4 are presented in **a**, **b**, **c**, and **d**.

The structure of BH and HO connections is presented in Figure 8a-d: the bias and output connections to atoms degrees 1 are presented in Figure 8a; those connecting the atoms with degrees 2, 3, and 4 are depicted in Figure 8b-d, respectively. The bias neuron has also a connection to the output neuron (BO connection).

Considering all connection types present in the case of *t*-butylcyclopropane the total number of adjustable weights is: 4 (IH connections) + 4 (BH connections) + 4 (HO connections) + 1 (BO connection) = 13. MolNet can contain more parameters, corresponding to bonding relationships that are not present in this example, but are found in one or more molecules from the calibration set. A certain connection type has the same weight in all molecules that contain it.

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. Any optimization algorithm can be used to determine an optimal set of connection types. In the prediction phase MolNet uses the weights determined in calibration to compute the investigated molecular property for molecules that were not present in the calibration set. If the set of molecules used in the prediction phase contains bonding relationships that are absent in the molecules used in the calibration phase these bonding relationships are neglected in predicting the molecular property.

Using a vector of input data computed from the molecular graph, we present the signal flow through MolNet. The degree of an atom i from a molecular graph G, $\mathbf{DEG}_i = \mathbf{DEG}_i(G)$, is computed with the formula:

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

where $\mathbf{A} = \mathbf{A}(G)$ is the adjacency matrix. The degree vector of *t*-butylcyclopropane **2** is $\mathbf{DEG}(\mathbf{2}) = \{3, 2, 2, 4, 1, 1, 1\}$. After the generation of MolNet as presented in Figures 7 and 8, the vector of atomic descriptors is entered into the network through the neurons in the input layer. As is usual with neural networks, the input and output values are scaled. Each input neuron receives a numerical value representing the degree of the atom with the same label from the molecular graph. In the case of *t*-butylcyclopropane, input neuron 1 receives the value 3, neuron 2 receives the value 2, neuron 3 receives the value 2, neuron 4 receives the value 4, and neurons 5, 6, and 7 all receive the same value 1. In this way the **DEG** vector is entered into MolNet and is then propagated through the network following the rules of MLF ANN. The signal that is offered by the output neuron represents the computed value of the investigated property for *t*-butylcyclopropane.

5. CONCLUSIONS

Usually, a multi-layer feedforward neural network has a topology that does not depend on the structure of the chemical compounds used in calibration or prediction. The network receives information about the structure of the compounds only from the atomic or molecular descriptors received by the neurons in the input layer. In this review we have presented three new MLF neural models that encode the molecular structure into the topology of the neural network: ChemNet defined by Kireev, the neural device introduced by Baskin, Palyulin and Zefirov, and MolNet introduced by Ivanciuc. These neural models are developed specifically for chemical applications; using a set of rules each network encodes into its topology the structure of the molecule examined by the ANN. Depending on the structure of each molecule, the network changes the number of neurons in the input and hidden layers, together with the number and type of connections. Such neural models that use information on the chemical structure to generate the network were applied with success to QSPR studies.

REFERENCES

(1) Hopfield, J.J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U.S.A*. **1982**, *79*, 2554-2558.
(2) Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533-536.
(3) Rumelhart, D.E.; McClelland, J.L. *Parallel Distributed Processing*; MIT Press: Cambridge, MA, USA, 1986.
(4) Wasserman, P.D. *Neural Computing*; Van Nostrand Reinhold: New York, 1989.
(5) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Anal. Chim. Acta* **1991**, *248*, 1-30.
(6) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl*. **1993**, *32*, 503-527.
(7) Burns, J.A.; Whitesides, G.M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev*. **1993**, *93*, 2583-2601.
(8) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH: Weinheim, 1993.
(9) Bulsari, A.B. *Neural Networks for Chemical Engineers*; Elsevier: Amsterdam, 1995.
(10) Devillers, J. (1996). *Neural Networks in QSAR and Drug Design*. Academic Press, London, p. 279.
(11) Ivanciuc, O. Artificial Neural Networks Applications. Part 5. Prediction of the Solubility of C_{60} in Organic Solvents. *Rev. Roum. Chim*. **1998**, *43*, 775-783.
(12) Ivanciuc, O. Artificial Neural Networks Applications. Part 6. Use of Non-Bonded van der Waals and Electrostatic Intramolecular Energies in the Estimation of ^{13}C-NMR Chemical Shifts in Saturated Hydrocarbons. *Rev. Roum. Chim*., **1995**, *40*, 1093-1101.
(13) Ivanciuc, O. Artificial Neural Networks Applications. Part 8. The Influence of the Activation Function in the Estimation of the Sensorial Scores of Red Wine Color. *Rev. Roum. Chim*. **1998**, *43*, 545-551.
(114) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.P. ^{13}C NMR Chemical Shift Prediction of sp^2 Carbon Atoms in Acyclic Alkenes using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644-653.
(15) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.P. ^{13}C NMR Chemical Shift Prediction of the sp^3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 587-598.
(16) Ivanciuc, O. Molecular Graph Descriptors Used in Neural Network Models. In: *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A.T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999, pp. 697-777.
(17) Ivanciuc, O.; Balaban, A.T. Graph Theory in Chemistry. In: *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998, pp. 1169-1190.

(18) Ivanciuc, O.; Balaban, A.T. The Graph Description of Chemical Structures. In: *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A.T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999, pp. 59-167.
(19) Ivanciuc, O.; Ivanciuc, T.; Balaban, A.T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In: *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A.T., Eds.; Gordon and Breach Science Publishers: The Netherlands, 1999, pp. 169-220.
(20) Ivanciuc, O.; Ivanciuc, T.; Diudea, M.V. Molecular Graph Matrices and Derived Structural Descriptors. *SAR QSAR Environ. Res*. **1997**, *7*, 63-87.
(21) Diudea, M.V.; Gutman, I. Wiener-Type Topological Indices. *Croat. Chem. Acta* **1998**, *71*, 21-51.
(22) Gakh, A.A.; Gakh, E.G.; Sumpter, B.G.; Noid, D.W. Neural Network-Graph Theory Approach to the Prediction of the Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 832-839.
(23) Baskin, I.I.; Palyulin V.A.; Zefirov, N.S. A Neural Device for Searching Direct Correlations between Structures and Properties of Chemical Compounds. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 715-721.
(24) Kireev, D.B. ChemNet: A Novel Neural Network Based Method for Graph/Property Mapping. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 175-180.
(25) Ivanciuc, O. Artificial Neural Networks Applications. Part 9. MolNet Prediction of Alkane Boiling Points. *Rev. Roum. Chim*. **1998**, *43*, 885-894.
(26) Ivanciuc, O. The Neural Network MolNet Prediction of Alkane Enthalpies. *Anal. Chim. Acta* **1999**, *384*, 271-284.