Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes

Daniel Svozil,† Jiří Pospíchal, and Vladimír Kvasnička*

Department of Mathematics, Faculty of Chemical Technology, Slovak Technical University,
81237 Bratislava, Slovakia

Received April 10, 1995®

Three-layer feed-forward neural networks for the prediction of ^{13}C NMR chemical shifts of alkanes through nine carbon atoms are used. Carbon atoms in alkanes are determined by 13 descriptors that correspond to the so-called embedding frequencies of rooted subtrees. These descriptors are equal to numbers of appearance of smaller structural skeletons composed of two through five carbon atoms. It is demonstrated that the used descriptors offer a very useful formal tool for the proper and adequate description of environment of carbon atoms in alkanes. Neural networks with different numbers of hidden neurons have been examined. Best results are given by the neural network composed of three hidden neurons. Simultaneous calculations carried out by the standard linear regression analysis are compared with our neural network calculations.

INTRODUCTION

One of the most important properties of the so-called feed-forward neural networks[1] is their ability to work as a universal approximator[2] of functions determined by a table (or set) of values of independent and dependent variables. Recently, neural networks have frequently been used in chemistry[3,4] as mathematical systems for classification or pattern recognition of data (e.g., does this data set of structural features suggest a specified type of chemical or biological property or activity?). They are especially appropriate[4] when the relationship between input and output data is complex, rapid processing information is called for, and only approximate solution is needed.

In our recent papers[5,6] the so-called embedding frequencies[7−9] have been successfully used for prediction of some thermochemical properties of alkanes (boiling points, melting points, molar volumes, etc.) by neural networks. The purpose of this paper is to demonstrate that embedding frequencies are very useful as descriptors for local environment of carbon atoms in alkanes (in general, acyclic molecules). This observation should not be surprising for those chemists who believe that local properties of atoms in saturated organic acyclic molecules are well determined by their local environment, since embedding frequencies characterize this environment as numbers of appearance of some preselected smaller carbon skeletons. ^{13}C NMR chemical shifts of alkanes are a good example of local property mainly determined by its neighborhood molecule.

FEED-FORWARD NEURAL NETWORKS

The feed-forward neural networks[1] belong to basic types of neural networks and are probably the most widely used ones. Our description shall therefore give only the basic features of the used network, so that it can be reproduced by an experienced user of neural nets.

The used neural network had three layers, the input one with 13 neurons, the hidden one with 1−5 neurons (the number of hidden neurons was optimized), and the output layer with only one neuron. Each neuron in one layer was connected with all neurons in the next level. The output value (activity) of the output neuron described the normalized NMR chemical shift of the ^{13}C atom. The outline of the used network is shown in Figure 1.

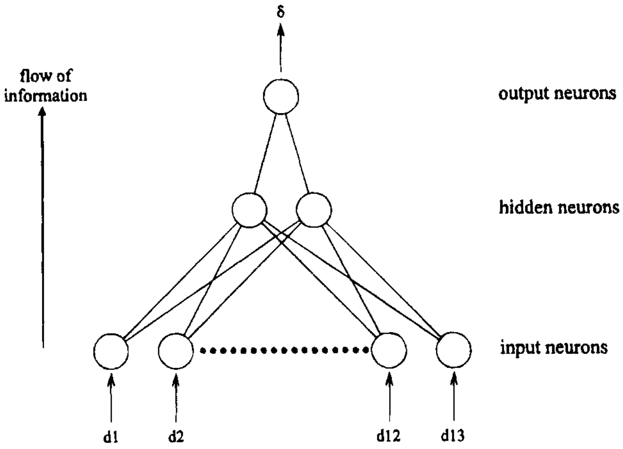

Figure 1. Structure of a three-layer neural network composed of input, hidden, and output layer of neurons. The layer of input neurons is composed of 13 neurons; their input activities are equal to descriptors d_1, d_2, ..., d_{13} that in turn are equal to the embedding frequencies of rooted subtrees (see Figure 2). The output layer is composed of one neuron, its activity being equal to $\delta/100$, where δ is ^{13}C NMR chemical shift of the given atom in the given alkane. The number of neurons in the hidden layer is varied from 1 to 5.

The activities x_i of the hidden and output neurons indexed i are determined by

$$x_i = f(\xi_i) \tag{1}$$

$$\xi_i = \vartheta_i + \sum_j \omega_{ji} x_j \tag{2}$$

where ϑ_i is the *threshold coefficient* assigned to the ith neuron and ω_{ji} is the *weight coefficient* of the connection from the jth neuron to the ith neuron. The summation runs over all neurons transferring signal to the ith neuron. The *transfer function* is

$$f(\xi) = \frac{1}{1 + \exp(-\xi)} \tag{3}$$

† Permanent address: Department of Analytical Chemistry, Faculty of Science, Charles University, Prague, Czech Republic.

® Abstract published in *Advance ACS Abstracts*, July 15, 1995.

0095-2338/95/1635-0924$09.00/0 © 1995 American Chemical Society

The *supervised adaptation process* changes the threshold and weight coefficients so that the sum of squared differences between the computed NMR shift and the measured NMR shifts is minimal. This is expressed by minimization of the *objective function E*

$$E = \frac{1}{2}(x_o - \hat{x}_o)^2 \qquad (4)$$

where x_o and $\hat{x}_o$ are computed and required activities of the output neuron, respectively.

The steepest-descent minimization method is accelerated by the so-called momentum method. Updating of weight and threshold coefficients was done by the following formulas

$$\omega_{ji}^{(k+1)} = \omega_{ji}^{(k)} - \lambda\left(\frac{\partial E}{\partial \omega_{ji}}\right)^{(k+1)} + \alpha\Delta\omega_{ji}^{(k)} \qquad (5a)$$

$$\vartheta_{j}^{(k+1)} = \vartheta_{j}^{(k)} - \lambda\left(\frac{\partial E}{\partial \vartheta_{j}}\right)^{(k+1)} + \alpha\Delta\vartheta_{j}^{(k)} \qquad (5b)$$

where λ is the rate of learning, which is set to a small positive number 0.04 and further optimized during calculation (when objective function in the current iteration was greater than in the previous iteration, the λ was decreased by 9%). The parameter α is the momentum parameter, set to 0.5. The numerical values of the parameters α and λ were chosen from preliminary experimental calculations, so that good convergence of the adaptation process was achieved. The index $k + 1$ corresponds to the currently changed value, the index k corresponds to the previous iteration, and $\Delta\omega$ and $\Delta\vartheta$ correspond to changes performed at the end of previous iteration. The partial derivatives are calculated by the back-propagation algorithm,[1] and the values in parentheses correspond to sums of the derivatives over all compounds in the training set. Initial values of weight and threshold coefficients for the first iteration have been randomly generated from the interval (−1,1). The adaptation was stopped after 10 000 iterations. We have experimentally verified that up to 10 000 steps the testing set error is monotonously decreasing. If more learning steps are used, then this error increases, which indicates overtraining of the neural network. The used topology of neural networks corresponds to the three-layer neural network with variable number of hidden neurons, from one to five hidden neurons, see Figure 1.

DESCRIPTORS

^{13}C NMR chemical shifts belong to the so-called local molecular properties, where it is possible to assign unambiguously the given property to an atom (vertex) of structural formula (molecular graph). In order to correlate ^{13}C NMR chemical shifts with the molecular structure we have to possess information about the environment of the given vertex. The chosen atom plays a role of the so-called root,[10] a vertex distinguished from other vertices of the molecular graph. For alkanes (acyclic saturated hydrocarbons that are classified graph-theoretically as trees),[10] embedding frequencies[7-9] specify the number of appearances of smaller rooted subtrees that are attached to the root of the given tree (alkane), see Figures 2, 3, and 4. In our recent publication[7] we have elaborated simple theory of embedding frequencies of rooted trees, and moreover we have invented simple depth-first search algorithm for the construction of embedding frequencies of rooted trees. In this publication we present an extensive table of embedding frequencies of all rooted subtrees through five vertices for all trees through eight vertices. Then each atom (a nonequivalent vertex in the tree) in an alkane (tree) is determined by 13 descriptors $\boldsymbol{d} = (d_1,d_2,...,d_{13})$ that are used as input activities of neural networks. The entry d_i determines the embedding frequency of the ith rooted subtree (see Figure 2) for the given rooted tree (the root is specified by that carbon atom of which the chemical shift is calculated). Their number and form are

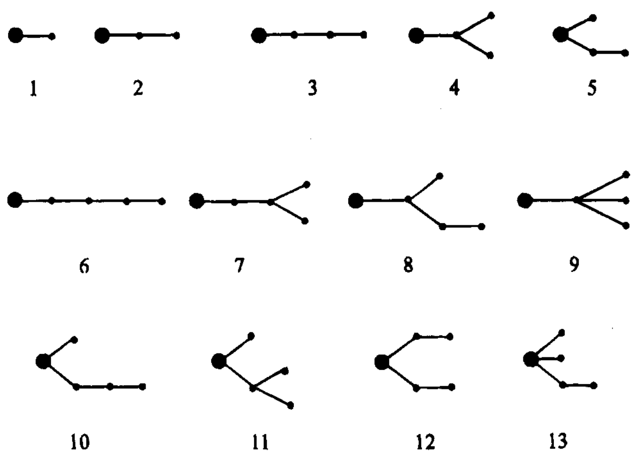

Figure 2. List of 13 rooted subtrees that are used for the calculation of embedding frequencies.

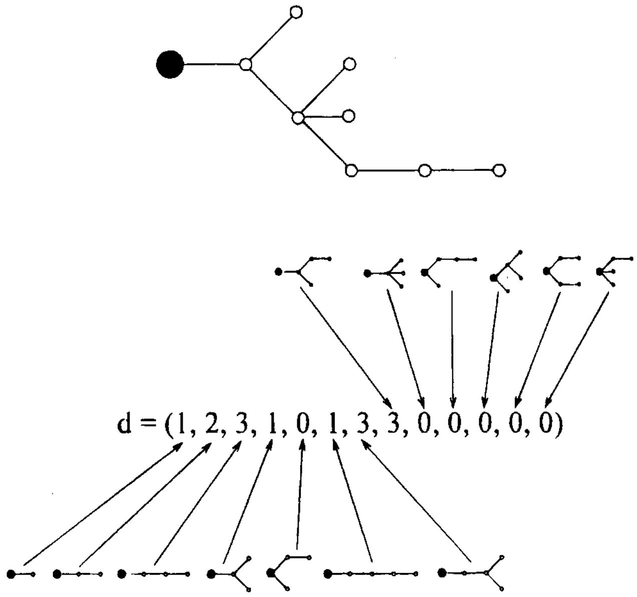

Figure 3. Illustrative example of embedding frequencies of a rooted tree (upper row); single entries of descriptor vector are assigned directly to their corresponding rooted subtrees.

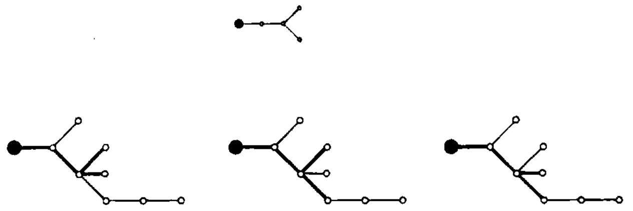

Figure 4. Illustrative example for a better understanding of embedding frequencies for the rooted tree displayed in Figure 3. For those given (upper row) there are displayed (lower row) single events when the rooted subtree is present in the rooted tree. The number of these events is equal to the embedding frequency.

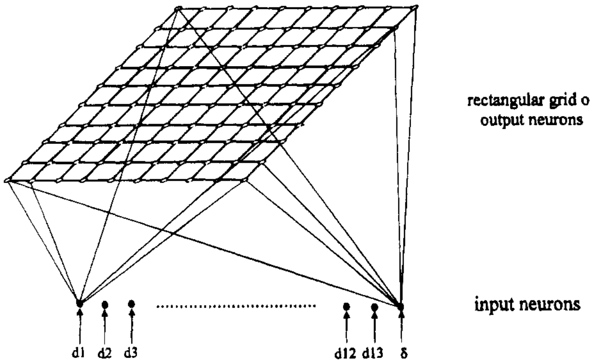

Figure 5. Schematic outline of Kohonen network composed of two layers. The upper layer contains output neurons situated at a rectangular grid so that each output neuron has two or three (marginal neurons) or four neighbors (inner neurons). The lower layer is composed of input neurons with input activities assigned to descriptors and the chemical shift, respectively. Each input neuron is connected with each output neuron; connections are evaluated by weights that are randomly initialized from open interval (0,1).

determined by our requirement to have all the rooted trees through five vertices. To avoid information redundancy, we have deleted those rooted trees, which embedding frequencies can be exactly determined from embedding frequencies of simpler rooted subtrees. This means that we consider at most δ-carbon effects. Recently, in neural network calculations as descriptors are very popular with some electronic characteristics of molecules, e.g., HOMO and LUMO energies, charges, etc. (e.g., ref 16). This work is trying to show that these descriptors are not necessary for alkanes, and more simple descriptors can be effectively used. Moreover, the calculation of electronic descriptors is comparable in complexity with neural networks, and if we know the wave function of a molecular system, then the chemical shifts can be calculated immediately by standard quantum chemical methods.

TRAINING AND TESTING SETS

^{13}C NMR chemical shifts of all alkanes from C_2 to C_9 available in the book[12] (cf. ref 11) (alkanes C_9 are not complete) are used as objects in our calculations. The total number of all alkanes considered in our calculations is 63; they give 326 different chemical shifts for topologically nonequivalent positions in alkanes. This set of 326 chemical shifts is divided into two disjoint subsets called the training set and the testing set. The training set is used for the adaptation process of neural networks, whereas the testing set is used for the verification of adapted neural networks.

The decomposition of a whole set of chemical shifts into training and testing sets was carried out by making use of the Kohonen neural network[13] with architecture specified by 14 input neurons and 15 × 15 = 275 output neurons situated on a rectangular grid 15 × 15, see Figure 5. The input activities of each object (chemical shift) are composed of 14 entries, whereby the first 13 entries are embedding frequencies, and the last, 14th entry, is equal to the chemical shift. Details of the used Kohonen network are described in Dayhoff's textbook[14] with parameters α = 0.2 (learning constant), d_0 = 10 (initial size of neighborhood), and T = 20 000 (number of learning steps). We have used the rectangular type of neighborhood, and the output activities were determined as L_1 (city-block) distances between input activities and the corresponding weights. After finishing the adaptation process, all 326 objects were clustered so that each object activates only one output neuron on the rectangular grid, and some output neurons are never activated and/or some output neurons are activated by one or more objects, see Figure 6. This means that this decomposition of objects through the grid of output neurons may be considered as a clustering of objects, each cluster, composed of one or more objects, being specified by a single output neuron. Finally, the training set is created so that we shift one object (with the lowest serial index) from each cluster to the training set and the remaining ones to the testing set. Then we get a training set composed of 112 objects and the testing set composed of 214 objects.

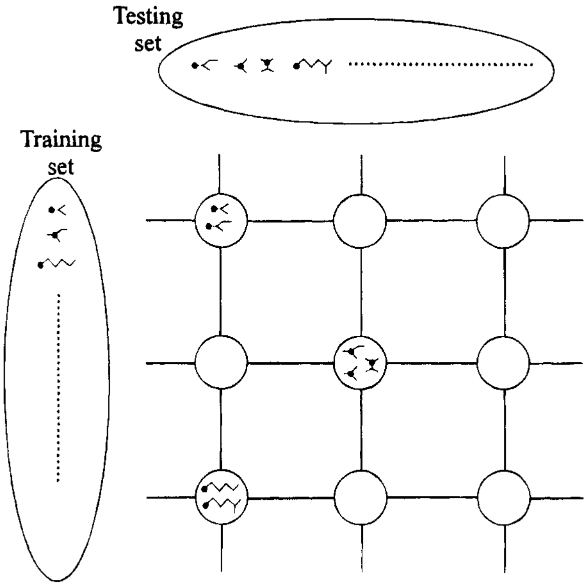

Figure 6. Kohonen network may serve as a clustering algorithm. After finishing its adaptation each object from the training set (a set composed of all 326 ^{13}C NMR shifts determined by 13 embedding frequencies and the value δ of chemical shift) activates just one output neuron. Neurons are schematically shown as circles, containing compounds (heavy dots specify carbon atoms for which the shifts are assigned) that activate the particular neuron. All compounds with particular chemical shifts that activate the same output neuron form a cluster. One representative (with smallest serial number) of each cluster is shifted to the training set (vertical ellipse), and the remaining ones (if any) are shifted to the testing set (horizontal ellipse).

Table 1. Results of Neural Network Calculations

type of neural network	training set		testing set	
	SEC	R^2	SEC	R^2
(13,1,1)	1.1387	0.9976	1.1913	0.9837
(13,2,1)	0.9906	0.9980	1.0980	0.9957
(13,3,1)	0.8941	0.9998	1.0732	0.9966
(13,4,1)	0.7517	0.9999	1.0905	0.9946
(13,5,1)	0.6656	1.0000	1.1041	0.9944

CALCULATIONS

The results of our neural network calculations for different numbers of hidden neurons (from one to five) are summarized in Table 1. The quantities SEC and R are determined as follows

$$SEC^2 = \frac{\sum (x_{\text{obs}} - x_{\text{calc}})^2}{N} \qquad (6)$$

Table 2. Results of LRA Calculations

type of LRA	training set		testing set	
	SEC	R^2	SEC	R^2
all objects[a]	0.9994	0.9900		
training set	0.9307	0.9893	1.1624	0.9872

[a] Training set is composed of all 326 objects.

$$R^2 = 1 - \frac{\sum(x_{\mathrm{obs}} - x_{\mathrm{calc}})^2}{\sum(x_{\mathrm{obs}} - x_{\mathrm{mean}})^2} \tag{7}$$

We see that the best results are produced by the neural network (13,3,1) composed of three hidden neurons, its SEC value for objects from the testing set being the lowest one. We can observe the following interesting property of feed-forward neural networks: The SEC value for the training set monotonously decreases when the number of hidden neuron increase; on the other hand, the SEC value for the testing set has a minimum for three hidden neurons. This means that the predictability of neural networks for testing objects is best for three hidden neurons, further increasing their number does not provide better results for testing set.

In the framework of linear regression analysis (LRA) chemical shifts (in ppm units) are determined as a linear combination of all 13 descriptors plus a constant term

$$\delta = c_0 + \sum_{i=1}^{13} c_i d_i \tag{8}$$

Two different LRA calculations have been carried out. While the first calculation was based on the whole set of 326 objects (chemical shifts), the second calculation included only the objects from the training set (the same as for neural network calculations). The obtained results are summarized in Table 2. Coefficients c_i of LRA (8) are listed in Table 3.

Comparing results of neural network and LRA calculations, we see that the best neural network calculation provides slightly better results for training objects than LRA. The SEC testing value for neural network calculation is slightly smaller than it is for LRA calculation. Table 4 lists precision of predictions of chemical shifts. It means, for instance, that the neural network (13,3,1) calculation for objects from the testing set (eighth column in Table 4) provides the following prediction: for 74% (78% and 88%) of the shifts, the difference between the experimental and predicted values was less than 1.0 ppm (1.5 and 2.0 ppm, respectively). On the other hand, what is very surprising, the LRA based on the training set gave a slightly better prediction for testing objects than the neural network (13,3,1) calculation. Precision of predictions for differences 1.5 and 2.0 ppm were slightly greater for LRA than for NN (neural network), see the sixth and eighth columns in Table 4.

CONCLUSIONS

Embedding frequencies of rooted trees are useful descriptors for description of environments of specific carbon atoms in alkanes. They provide graph-theoretically well defined[7-9] entities easily generalized also for acyclic organic molecules with heteroatoms (cf. >ref 9). Linear regression analysis and neural network calculations have been useful in developing mathematical models that predict ^{13}C NMR chemical shifts in alkanes through nine carbon atoms from embedding frequencies. It was demonstrated that both these methods produce results that are roughly comparable in their ability to predict ^{13}C NMR chemical shifts in alkanes. This is caused by the fact that embedding frequencies adequately describe the neighborhood environment of carbon atoms in alkanes so that the even unsophisticated LRA already provides accurate predictions. This means that possible nonlinearities in the relationship between embedding frequencies and chemical shifts are of small importance. An effectiveness of neural network calculations results from the fact that nonlinearities of input−output relationships are automatically taken into account. Since, as was mentioned above, nonlinearities in relationships between embedding frequencies and ^{13}C NMR chemical shifts in alkanes are of small (or negligible) importance, neural network calculations could not provide considerably better results than LRA calculations. Finally, as a byproduct of our LRA calculations, we have obtained simple linear relationships between ^{13}C NMR chemical shifts in alkanes and embedding frequencies which are more precise (see Table 4) than similar

Table 3. Coefficients of LRA

type of LRA	coefficients						
	c_0	c_1	c_2	c_3	c_4	c_5	c_6
all objects	−3.1292	8.5318	12.178	−3.8592	−4.0609	−1.9513	0.5222
training set	−3.4674	8.7676	11.823	−3.3845	−3.9009	−1.8696	0.7315

type of LRA	coefficients						
	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}
all objects	0.5186	0.3233	1.7001	0.9103	0.9019	0.0200	−1.5262
training set	0.3673	0.1094	1.9835	0.7551	0.8814	0.1315	−1.6782

Table 4. Precision of Prediction[a]

prediction precision (ppm)	Grant ref 15 (%)	Lindeman ref 11 (%)	LRA[b] all objects (%)	LRA[c]		NN (13,3,1)	
				training (%)	testing (%)	training (%)	testing (%)
1.0	61	61	78	78	69	87	74
1.5	77	78	89	90	85	96	78
2.0	84	89	94	97	91	98	88

[a] Rows indicate percentages of objects predicted by the given model with precision specified by maximum ppm absolute error shown in the first column. [b] LRA which is used for all 326 objects for training set. [c] LRA which is used for only 112 objects for training set.

relationships constructed by Grant[15] or Lindeman[11] often used in literature (cf. ref 12).

ACKNOWLEDGMENT

This research was supported in part by 1994 Chemical Structure Association Trust Award (UK).

REFERENCES AND NOTES

(1) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processes;* Rumelhart, D. E., McClelland, J. C., PDP Research Group, Eds.; MIT Press: Cambridge, MA, 1987; Vol. 1, pp 318–362.

(2) Hecht-Nielson, R. *Neurocomputing*; Addison-Wesley: Reading, MA, 1990.

(3) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists. An Introduction;* VCH: Weinheim, Germany, 1993.

(4) Burns, S. A.; Whitesides, G. M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev.* **1993**, *93*, 2583–2601.

(5) Cherqaoui, D.; Villemin, D.; Kvasnička, V. Application of Neural Network Approach for Prediction of Some Thermochemical Properties of Alkanes. *Chemometrics Intell. Lab. Syst.* **1994**, *24*, 117–128.

(6) Cherqaoui, D.; Villemin, D.; Mesbah, A.; Cence, J. M.; Kvasnička, V. *J.* Use of a Neural Network to determine the Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals and Their Sulphur Analogues. In *Chem. Soc., Faraday Trans.* **1994**, *90*, 2015–2019.

(7) Poshusta, R. D.; McHughes, M. C. Embedding Frequencies of Trees. *J. Math. Chem.* **1989**, *3*, 193–215.

(8) McHughes, M. C.; Poshusta, R. D. Graph-theoretic Cluster Expansion. Thermochemical Properties for Alkanes. *J. Math. Chem.* **1990**, *4*, 227–249.

(9) Kvasnicka, V.; Pospichal, J. Simple Construction of Embedding Frequencies of Trees and Rooted Trees. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 121–128.

(10) Harary, F. *Graph Theory;* Addison Wesley: Reading, MA, 1969.

(11) Lindeman, L. P.; Adams, J. Q. Carbon-13 Nuclear Magnetic Resonance Spectroscopy: Chemical Shifts for the Paraffins though C_9. *Anal. Chem.* **1971**, *43*, 1245–1252.

(12) Kalinowski, H. O.; Berger, S.; Braum, S. *^{13}C NMR Spektroskopie;* G. Thieme Verlag: Stuttgart, 1984.

(13) Kohonen, T. *Self-Organization and Associative Memory*; Springer Verlag: Berlin, 1988.

(14) Dayhoff, J. *Neural Network Architectures;* Van Nostrand Reinhold: New York, 1990.

(15) Grant, D. M.; Paul, E. G. Carbon-13 Magnetic Resonance II. Chemical Shift Data for the Alkanes. *J. Am. Chem. Soc.* **1964**, *86*, 2984–2990.

(16) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 68–76.

CI9500389