Neural Networks and ^{13}C NMR Shift Prediction

J. P. Doucet,* A. Panaye, E. Feuilleaubois, and P. Ladd

Institut de Topologie et de Dynamique des Systèmes associé au CNRS, URA-34, Université Paris 7, 1 rue Guy-de-la-Brosse, 75005 Paris, France

Received July 28, 1992

Computational neural networks are known to have the capability to predict complex mappings between input and output data. These new tools seem to be well-suited to NMR data. To treat simultaneously a whole set of compounds in the alkane family, we used a back-propagation neural network with a topological description as input. The results allow for a good prediction of the shifts because of the range of the test population (up to 62% of the known environments) and since all types of carbons are taken into account without distinction of connectivity.

INTRODUCTION

^{13}C nuclear magnetic resonance (^{13}C NMR) is largely used in structural analysis, taking advantage of the heavy dependence of ^{13}C chemical shifts on the structural environment of the resonating carbons.[1-3] Particularly, it plays a key role in spectral simulation (reconstruction of the spectrum associated to a given structural formula), which intervenes in conformity assessment of chemical samples (verifying a structural formula) and constitutes an important part of structural elucidation programs as a screening step to rank the proposed candidates. Spectral simulation can procede by the recognition of relevant fragments to which characteristic spectral features (here the chemical shifts in ^{13}C NMR) are associated. However, it more frequently relies on models or relationships trying to predict the ^{13}C shifts from structural characteristics; an approach faced by the conflicting requirements of a large structural scope and a good spectral precision.

Up to now, the mechanisms acting on the ^{13}C shifts are not quite well-understood, and shift prediction is largely carried out by means of empirical models relating structural descriptors to ^{13}C shifts in additive relationships[4,5] or linear correlations.[6] Current developments are associated with the introduction of statistical techniques such as factorial analyses,[7-9] and more recently neural networks offer new perspectives in this field because they can automatically detect important features in a data set and can provide a nonlinear mapping scheme.[10,11]

The aim of this paper is, therefore, to investigate the ability of neural networks to predict ^{13}C shifts with an accuracy convenient to spectral simulation. The acyclic alkanes were chosen as a reference population for this evaluation. Beside the fact that a large number of data are available from literature,[1,4,5] this chemical family offers widely varied carbon frameworks, and so they constitute a good example to examine whether branching effects (a source of difficulty in the previous predictive approaches) can be taken into account. Alkanes appear also as a privileged and largely used set of compounds for testing the capabilities of various structural descriptors (and among them diverse topological indices) in the prediction of physicochemical properties (boiling point, heat of formation ...). During the progress of this work, a paper from Anker and Jurs[10] appeared, tackling a similar approach for the ^{13}C shift prediction on ketosteroid carbons. The structural descriptors used involved atom–atom distances, van der Waals energies, and electronic charges; whereas, in our work we prefer a simpler topological description. Application of neural networks to the ^{13}C shifts calculation of alkanes was also presented by Kvasnicka,[11] but the topological descriptors used imply both some redundancy and nonunivocity, and the limited set studied does not allow for a clear perception of the predictive capability of the approach on a large family bearing heavily branched compounds.

Numerous and diverse approaches tackle the problem of ^{13}C shifts prediction. Pioneering works from Grant and Paul and others[1,4,5] used parametric additive models. The structural scope covered can be large, making it possible to use these shift increments in simulation programs but with limited precision only.[2,3] Extension to polysubstituted compounds or to more branched structures as in the Lindeman and Adams model[5] needed the adjunction of numerous interaction terms, and the overall precision remain limited (about 0.8 ppm in the model of Lindeman and Adams[5]). Within restricted populations of homogeneous compounds, better precision can be achieved, at the expense of structural generality, by linear regression models using either behavior/behavior (δ/δ) relationships[12] or correlations with structural descriptors.[6] An alternative approach was offered by topological correlations: In the DARC PELCO method, the environment of the resonating carbon is concentrically described in terms of discrete and ordered atom sites which are given individual contributions to the chemical shift (these values being derived from least-squares correlations on a selected learning set: the "key population"). For a given compound, the property (here the ^{13}C shift) is then evaluated by the summation of the contributions associated to the sites which are occupied in the environment of the resonating carbon and some interaction terms.[13,14] These topological correlations are able to give very precise predictions for the ^{13}C shifts, and the individualization of site influences is very powerful to detect nonadditive behaviors, i.e., interactions between sites when they are simultaneously occupied. However, various examples showed that such topological correlations give good results only if the connectivity of the resonating carbon remains constant, leading to split chemical families into subsets for primary, secondary, ..., quaternary carbons.

One of our objectives was, therefore, to see if the shrewdness of the topological descriptors can be incorporated for spectral simulation purposes into a more global model, gathering carbons of varied connectivity.

Apart from these topological models, the approach of Beierbeck and Saunders[15] would, of course, deserve special attention since it takes into account the spatial organization of the environment. However, as it requires weighted averages on various conformers, it is not so easily incorporated into simulation systems.

Experimental Section. The chemical shifts used in this study were taken from Lindeman and Adams[5] for alkanes from C5

0095-2338/93/1633-0320$04.00/0

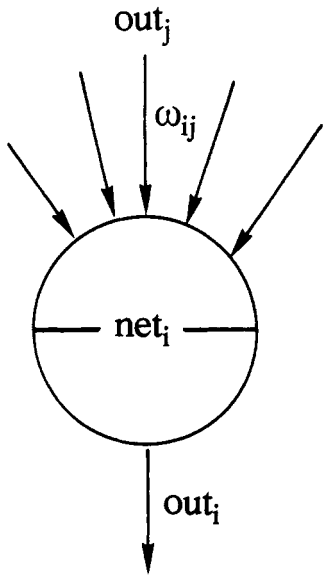


Figure 1. Formal neuron.

to C9, and the shifts for the first compounds were taken from ref 1. Chemical shifts values are relative to TMS (tetramethylsilane). Since the sigmoidal transfer function, used in the neural network treatment, only provides outputs in the range [0,1] these shifts were scaled according to

$$\text{target} = (\delta + 2.3)/59.1$$

Neural network simulation was carried out, thanks to the algorithm of McClelland and Rumelhart,[16,17] on a Digital Equipment workstation DECSTATION 5000/200. Training the network (in its 27/6/1 units configuration) typically requires 120 s for running 3000 epochs on a learning set of about 100 compounds.

NEURAL NETWORKS

Recent papers and reviews detail the mode of functioning of artificial neural networks.[10,11,16-18] Neural networks are formed by simple processing units (neurons) interconnected between themselves (network). In a feed forward layered architecture (like the one used here), each neuron receives weighted inputs from each neuron in the preceding layer. From the sum of these values (net input), it calculates an output value by means of a transfer function (generally the sigmoid function) and transmits that value to the neurons of the following layer:

$$\text{net}_i = \sum \omega_{ij}\text{out}_j + \theta_i$$

$$\text{out}_i = 1/[1 + \exp(-\text{net}_i)]$$

where net$_i$ is the net input of neuron i; θ_i is the bias of neuron i; and ω_{ij} is the weight of the connection between upstream neuron j and neuron i (see Figure 1).

The architecture chosen here consists of an input layer (accepting the topological parameters defining the ^{13}C environment), a hidden layer ensuring the encoding of the interactions in the data, and an output layer that delivers the calculated chemical shift. (See Figure 2).

The network operates in supervised process: In a first (training) phase, the network is fed with a set of known input (topology)/target (chemical shift) couples forming the training set. The connection weights and the biases are iteratively modified (from initial random values) so that the network tends to reproduce as outputs the target values (here the chemical shifts). This adjustment is carried out according to

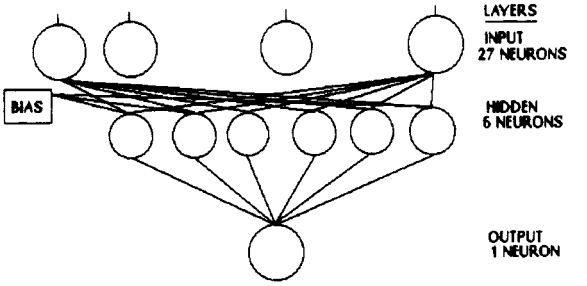


Figure 2. Typical architecture of a feed forward artificial neural network (input layer, hidden layer(s), output layer). No bias was added to the output unit.

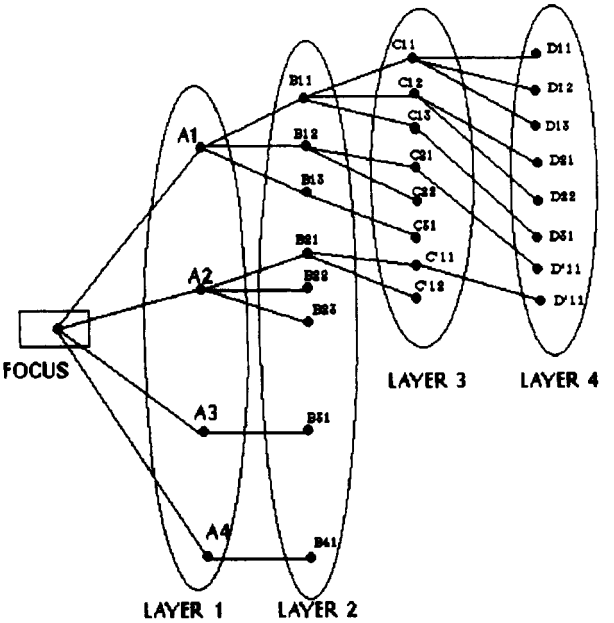
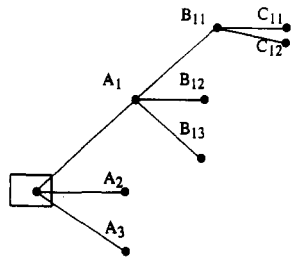


Figure 3. Topological description of the studied population.

the algorithm of back-propagation of errors by minimization of the sum of squared differences between known target values and calculated output values, weights adjustments processing backward from the output layer toward the input layer. After this training phase, connections and biases are frozen, and the network is now ready to calculate as output an (unknown) δ value for a carbon whose environment topological descriptor is given as input.

In this work, the input layer is constituted by a topological description of the environment of the resonating carbon according to the concepts of the DARC system. Starting from the resonating carbon chosen as focus, the sites of the environment occupied by carbons are described in concentric ordered ranks, A, B, C, ..., (Figure 3). Within each rank, an order is imposed on the occupied sites. An application example of these topological descriptors is given on Figure 4 (position B_{23} can only be occupied if position B_{22} and the preceding ones are occupied). Four ranks of atoms were considered since it is well-accepted that substituents farther apart than δ position have nearly no influence on the ^{13}C shifts. From the trace of the population studied (208 differing environments in 65 compounds), schematized in the graph of Figure 3, it appears that all the environments considered can be represented by a string of 27 (ordered) binary values (1 if the corresponding position is occupied, 0 otherwise) characterizing univocally each individual environment and constituting the input values. The 27 neurons of that input layer are connected to a hidden layer for which we will optimize the number of units (see below). These hidden units are connected to one output neuron giving the chemical shift.



A1	A2	A3	A4	B11	B12	B13	B21	B22	B23	B31	B41	C11	C12	C13	C21
1	1	1	0	1	1	1	0	0	0	0	0	1	1	0	0

C22	C31	C'11	C'12	D11	D12	D13	D21	D22	D31	D'11
0	0	0	0	0	0	0	0	0	0	0

Figure 4. Ordered topological descriptor.

OPTIMIZING ARCHITECTURE AND PARAMETERS OF THE NETWORK

Starting from the whole population (208 couples environment/δ), we first determine the best number of hidden units. This was carried out iteratively: In the current training step, for a given number of hidden units, we see if convergence is reached (that is if the network stabilizes its connections and biases with an acceptable output error), and if so, the next training phase is run with one unit less in the hidden layer ... and so on until convergence is not achieved. Starting arbitrarily from a hidden layer of 10 units, we observed that the hidden layer can be reduced from 10 to 6 units without significantly lessening the data restitution. For example, results obtained with 6 or 7 hidden units are quite similar as to the convergence rate and the data restitution: the sum of the squared differences between target and calculated outputs, tss, is about 0.011 after 5000 epochs for the whole set of 208 scaled δ values. But with 5 hidden units, stabilization of the network leads only to a tss of about 0.020 after the same number of epochs, and this value does not decrease even if many more iterations are carried out, in contrast with the configuration with 6 or 7 hidden units where the tss continuously decreases when more iterations are run. In this exploratory work, no attempts were carried out to use more than one hidden layer. It may be hoped that a multi-hidden-layer organization would reduce the number of weights to stabilize, and the training time, but it would surely make the interpretation of the results more difficult because of the more diffuse character of the information. So we retained the following architecture: 27 input neurons, 6 hidden neurons, and 1 output neuron.

Similarly we optimize the two parameters intervening in the correction of errors in the back-propagation algorithm:

$$\Delta\omega^n{}_{ij} = \epsilon\delta^n{}_i \mathrm{out}^{n-1} + \mu\Delta\omega^n{}_{ij}{}^{(\mathrm{prec.})}$$

For neuron i of the current layer n the learning rate, ϵ, is a proportionality constant which monitors the changes in the connection weights, and the momentum, μ, indicates what ratio of the preceding adjustment step ($\Delta\omega^n{}_{ij}{}^{(\mathrm{prec.})}$) will be retained in the current step.

Two series of the test have been carried out to determine the best values of these parameters, which at different levels monitor the velocity of the network to learn and to stabilize while avoiding random oscillations which slow the convergence.

First, for a given momentum (0.9, a commonly used value), various learning rates were tried from 0.05 to 0.15 on runs of 3000 epochs. Convergence is always achieved, but for the lower and upper values, some instability is observed during the training with sudden rises of the total error (tss) and come back to values near to equilibrium, prompting us to retain a learning rate of 0.1.

Then, for a fixed learning rate of 0.1, various values from 0.7 to 0.95 were used for the momentum. Here also the extreme values lead to oscillations, and 0.9 was retained.

RESULTS

Six analyses to test the capability of the network to reproduce or predict the alkane carbons' chemical shifts, depending on the nature of the learning and test population, have been carried out. Statistical criteria are gathered in Table I (analyses are numbered from I to VI). They will be briefly discussed below.

Correlation. A first application uses the whole set of data available (208 topological environments corresponding to different shift values). The mean error obtained after stabilization of the best network (0.32 ppm), with only 0.96% shifts calculated farther than 1 ppm from experiment, indicates that the network is quite able to adapt its connections to the submitted (input/output) couples and to recall data with a quite satisfactory precision (an accuracy of ca. 0.5 ppm seems quite convenient to encompass the usual medium effects on ^{13}C shifts). It can be noted also that the widely used model of Lindeman and Adams[5] on the same population corresponds to a standard error of 0.79 ppm, with 15% shift deviation larger than 1 ppm.

These encouraging results prompt us to examine the capability of the network for prediction: evaluation of the chemical shift for compounds not previously included in the learning set.

Prediction. For prediction applications, the network is first trained on a learning set comprising only a part of the total set of environments. Then, freezing the weights obtained, it is used to calculate the chemical shifts for other compounds constituting a test set. Comparing these calculated results with the experimental values allows us to testify of the ability of the network to acquire its knowledge for prediction of chemical shift of environments not presented to the network during the training phase. Various stages have been examined depending on the relative importance of the learning and test sets.

For such applications, overtraining may constitute an obstacle if the network has many degrees of freedom (adjustable weights and biases): when more and more epochs are run, the network can organize itself to map better and better the learning set. But this learning "by heart" makes the network lose its capability of generalization: unknown situations are badly treated. With a growing number of iterations, the error on the learning set continuously decreases, that on the test set first decreases and then increases when overtraining occurs. To prevent such problems, calculations on the test set were performed at regular intervals of the learning phase so as to get the optimal prediction on the test set.

In a first example, the test set was extracted from the whole population by a random selection of 30 environment values. After training on the remaining 208 – 30 = 178 values, the network is able to calculate the 30 test shifts with a rms deviation of 0.54 ppm (leaving only 6.7% shifts farther than 1 ppm from the experimental values).

Our aim being the prediction of shifts, efforts were subsequently made in the following studies to enlarge the test set and limit the learning set. Since this learning set must have some information about all the possible occupied sites, we chose to select for the learning set the minimal environment of each site, for every connectivity (1–4) of the resonating

Table I. Analysis of the Results[a]

analysis	I	II	III	IV	V	VI
			Learning Set			
no. of patterns	208	178	81	77	70	65
range	−2.3 to +56.8	−2.3 to +56.8	−2.3 to +56.8	−2.3 to +56.5	−2.3 to +56.5	−2.3 to +56.5
mean δ	28.3	28.2	27.9	27.8	27.6	26.2
rms deviation	0.3	0.4	1.6	1.3	1.4	0.9
			Test Set			
no. of patterns		30	127	131	138	143
range		+15.9 to +46.6	+7.5 to +51.0	+7.5 to +56.8	+7.5 to +56.8	+7.5 to +56.8
mean δ		29.1	28.6	28.6	28.7	29.3
rms deviation		0.6	1.0	1.3	1.6	1.7
mean error		0.5	0.8	1.0	1.2	1.3
no. of outliers, ppm						
>5	*b*	0	0	1	1	2
>3	0	0	1	4	10	11
>1	2	2	46	54	66	74
>0.5	22	15	77	89	99	108

[a] Range, mean δ (ppm); lower, upper, and mean δ values in the set. rms deviation, root mean-square deviation between calculated and experimental δ values. mean error, mean of the differences (absolute values) between calculated and experimental δ. [b] Full data set.

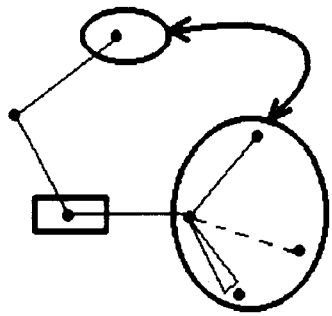


Figure 5. *tert*-Butyl interaction.

carbon (in the DARC description, the minimal environment of site *i* comprises the minimal set of sites necessarily occupied in the population studied for occupation of site *i*). This criterion defines a learning set of 65 compounds. After the training, the network calculates the shifts of the remaining 143 compounds (test population) with a mean error of 1.26 ppm and 74 compounds (51%) deviating by more than 1 ppm. This rather poor result can be due to the large size of the test set (68.7% of the total population) compared to that of the learning set (31.3%), but it more likely reflects the fact that only the connectivity of the ^{13}C was considered, so that the network is unable to learn other important interactions between sites in cases that were not presented to it in the training phase.

In the subsequent attempts, we proposed to include in the learning set some compounds representative of interactions between sites which are likely to intervene. The hope was that the network can organize its connections to take into account such situations and so become able to fairly evaluate them in unknowns (i.e., environments not included in the training set and submitted to the network for shift prediction). So, in the third example, because of large deviations (greater than 4 ppm) observed for some environments in the preceding run, we introduce in the learning set five environments bearing a *tert*-butyl interaction (see Figure 5), that is environments where a *tert*-butyl group is able to interact with other B carbon-(s). After the learning (now on 70 environments), the calculations carried out on the remaining 138 chemical shifts give a mean error of 1.21 ppm, and now 66 shifts are calculated with a deviation (by respect to the experimental value) greater than 1 ppm, in place of 74 in the preceding run. Once more, examination of the outliers shifts prompts us to consider in the learning set supplementary types of interactions involving triads of successive tertiary or quaternary carbons (Figure 6) for a secondary or tertiary focus, corresponding to congested

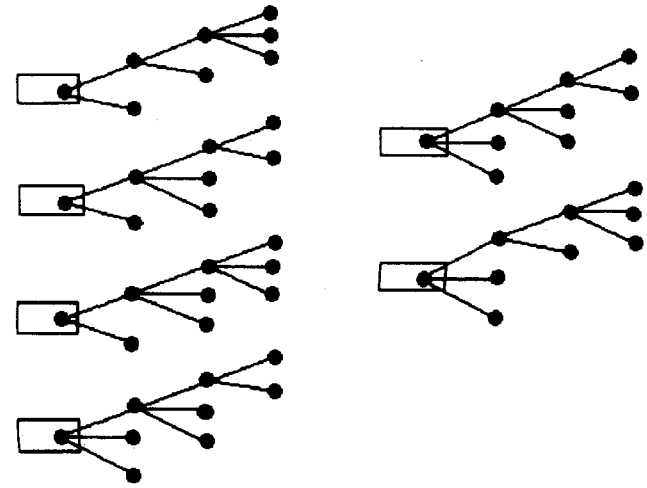


Figure 6. Triads of highly branched carbons.

environment. Adding seven new environments to the learning set allows for a mean error on the test set (131 shifts) of 1.0 ppm with 54 deviations only larger than 1 ppm. At last, including in the learning set, the maximal environments with C positions for each connectivity of the B carbons (four cases) leads to a mean error of only 0.83 ppm with 46 shifts only deviating by more than 1 ppm along 127 environments (the test set corresponds now to 61.1% of the available data).

CONCLUSION

As a conclusion it seems to us from this exploratory study that neural networks can be efficiently applied for the prediction of the ^{13}C shifts with acceptable precision and capability of generalization. Compared to various other models relying on topological descriptors, it is quite noteworthy that the network allows for prediction over the whole data set with no need to split the population into subsets of given connectivity.

However, some important features must be kept in mind:

Such networks offer a large number of parameters, the choice of which influences the results: selection of the network architecture, number of hidden units, value of learning rate, momentum, ...

A good selection of the learning set is also determinant. Careful examination of the results, allowing us to perceive what types of structural environments are not well-treated (or learned), may be of great help for improving predictions.

The possibility of local optimums where the learning algorithm may get stuck in must also be kept in mind owing to the random weights values in the starting configuration.

At last, the problem of the degrees of freedom in the network is to be considered in view of possible use for structure/property correlation (and derived applications in spectral simulation) by comparison with other approaches. The network used here has as much as 174 degrees of freedom (including all connections and biases). This number is quite large and even may be greater than the number of observations in the learning set for some cases we examined. This is not an uncommon situation in neural network applications. But an important consequence is that similar quality solutions can arise from largely different matrices of weights, the weights determined after the training phase being heavily dependent of their random starting values, so that any interpretation of the weights is not attainable in such conditions.

Of course it is not surprising that with a large number of degrees of freedom, the network can map a given data set with high accuracy (higher than that attained for usual models such as that of Lindeman and Adams, which require less parameters). But the important points are that:

(1) The network can organize its weights in such a way that it can calculate quite correctly the shifts for compounds not belonging to the training set, and so presents a real predictive ability,

(2) this adjustment is made automatically by the system itself, with no need of a predefined (linear, parabolic, ...) model.

These non-fully-deterministic characters are not a real drawback when only the prediction of unknown values from the knowledge extracted from known neighboring situations is sought for, as this is the aim of simulation applications. But of course, in such underdetermined systems, with many degrees of freedom, it is not easy to understand which task the network is actually performing. The situation is, therefore, quite different from that encountered in the widely-used linear regression models appearing in free energy relationships. In such models the slopes or the regression coefficients give some information about the physicochemical mechanisms involved, once the good structural parameters have been selected (i.e., those representing the real mechanisms, steric, inductive, resonance, ..., intervening, and constituting non-intercorrelated scales). However, despite these obvious difficulties for interpretative studies (at least for systems with too large degrees of freedom), neutral networks appear quite attractive for prediction in complex sets of data, without the need for a precise definition of a representation model, the system being able to organize itself for a good mapping after learning on a subset of known data.

Using a multi-hidden-layers network may constitute an attractive way which deserves special attention to reduce the number of degrees of freedom, but with the remaining difficulty of interpretating a more intricately distributed information. We plan to examine this point in the near future.

SUPPLEMENTARY MATERIAL

Table of the topological descriptors of alkane carbons and chemical shift values (5 pages). Ordering information is given on any current masthead page.

REFERENCES AND NOTES

(1) Kalinowski, H. O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*; John Wiley and Sons: New York, 1988.

(2) Cheng, H. N.; Bennett, M. A. Trends in Shift Rules in Carbon-13 Nuclear Magnetic Resonance Spectroscopy and Computer-Aided Shift Prediction. *Anal. Chim. Acta* **1991**, *242*, 43–56.

(3) Pretsch, E.; Fürst, A.; Robien, W. Parameter Set for the Prediction of the ^{13}C-NMR Chemical Shifts of sp^2- and sp-hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1991**, *248*, 415–428.

(4) Grant, D. M.; Paul, E. G. Carbon-13 Nuclear Magnetic Resonance. II. Chemical Shifts Data for the Alkanes. *J. Am. Chem. Soc.* **1964**, *86*, 2984–2989.

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

(6) Sutton, G. P.; Anker, L. S.; Jurs, P. C. Evaluation of Automated Methods for the Selection of Models for Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Keto Steroids. *Anal. Chem.* **1991**, *63*, 443–449, and references therein.

(7) Wiberg, K. B.; Pratt, W. E.; Bailey, W. F. Nature of Substituent Effects in Nuclear Magnetic Resonance Spectroscopy. 1. Factor Analysis of Carbon-13 Chemical Shifts in Aliphatic Halides. *J. Org. Chem.* **1980**, *45*, 4936–4947.

(8) Doucet, J. P.; Yuan, S. G.; Dubois, J. E. Evolvement of Alpha Functional Substituent Shifts in ^{13}C NMR: DARC PULFO Topological Correlation. *J. Chim. Phys.* **1984**, *81*, 219–224.

(9) Doucet, J. P.; Panaye, A.; Yuan, S. G.; Dubois, J. E. Evolvement of Alpha Functional Substituent Shifts in ^{13}C NMR: Application of the DARC PULFO Topological Model for Acyclic Derivatives. *J. Chim. Phys.* **1985**, *82*, 607–611.

(10) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157–1164.

(11) (a) Kvasnicka, V. An Application of Neutral Networks in Chemistry. *Chem. Pap.* **1990**, *44*, 775–792. (b) Ibid. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63–76.

(12) (a) Ejchart, A. Substituent Effects on ^{13}C NMR. 2-Chemical Shifts in the Saturated Framework of Secondary Aliphatic Derivatives. *Org. Magn. Reson.* **1981**, *15*, 22–24. (b) Substituent Effects on the ^{13}C NMR Chemical Shifts in the Saturated Framework of Primary Aliphatic Derivatives. *Org. Magn. Reson.* **1980**, *13*, 368–371.

(13) Dubois, J. E.; Doucet, J. P. ^{13}C NMR of Aliphatic Alkynes: Topolgical Analysis of Alkyl Substituent Effects on the Chemical Shift of sp Carbons by the DARC PELCO Method. *Org. Magn. Reson.* **1978**, *11*, 87–96.

(14) Dubois, J. E.; Doucet, J. P.; Tiffon, B. J. Carbon 13 NMR: Alkyl Substituent Effects on the Chemical Shift of the Carbonyl Carbon in Aliphatic Ketones. *J. Chim. Phys.* **1973**, 805–806.

(15) Beierbeck, H.; Saunders, J. K. Conformational and Configurational Analysis of Hydrocarbons Chains Based on Time-Averaged Carbon-13 Chemical Shifts. *Can. J. Chem.* **1980**, *58*, 1258.

(16) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*; MIT Press: Cambridge, MA, 1986.

(17) McClelland, J. L.; Rumelhart, D. E. *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*; MIT Press: Cambridge, MA, 1988.

(18) 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.