^{13}C NMR Chemical Shift Prediction of sp^2 Carbon Atoms in Acyclic Alkenes Using Neural Networks

O. Ivanciuc,[§] J.-P. Rabine,[†] D. Cabrol-Bass,[*,†] A. Panaye,[‡] and J. P. Doucet[‡]

Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest,
Splaiul Independentei 313, 77206 Bucharest, Romania, LARTIC University of Nice-Sophia Antipolis,
Parc Valrose 06108 Nice Cedex, France, and ITODYS, University of Paris 7 - Denis Diderot,
1 Rue Guy de la Brosse, 75005 Paris, France

Received September 14, 1995[⊗]

The ^{13}C NMR chemical shift of sp^2 carbon atoms in acyclic alkenes was estimated with multilayer feedforward artificial neural networks (ANNs) and multilinear regression (MLR), using as structural descriptors a vector made of 12 components encoding the environment of the resonating carbon atom. The neural network quantitative model provides better results than the MLR model calibrated with the same data. The predictive ability of both the ANN and MLR models was tested by the leave-20%-out (L20%O) cross-validation method, demonstrating the superior performance of the neural model. The number of neurons in the hidden layer was varied between 2 and 7, and three activation functions were tested in the neural model: the hyperbolic tangent or a bell-shaped function for the hidden layer and a linear or a hyperbolic tangent function for the output layer. All four combinations of activation functions give close results in the calibration of the ANN model, while for the prediction a linear output function performs better than a hyperbolic tangent one, but from a statistical point of view one could not choose a particular combination against the others. For the ANNs with four neurons in the hidden layer, the standard deviation for calibration ranges between 0.59 and 0.63 ppm, while for prediction it lies between 0.89 and 1.07 ppm. We propose a parallel use of the four ANNs for the prediction of unknown shifts, because the mean of the four predictions exhibit a smaller number of outliers with lower residuals. The present model is compared with three additive schemes for the calculation of the sp^2 ^{13}C NMR chemical shifts, and the statistical analysis of the results demonstrates that the ANN model gives better predictions than the classical ones.

1. INTRODUCTION

^{13}C NMR spectral simulation represents an important method for the identification of chemical compounds and for the validation of their spectral assignments. The methods used in predicting ^{13}C NMR shifts are diverse, ranging from database retrieval[1,2] to additive relationships[3−9] and empirical models which include various topological,[10] molecular, and quantum mechanics descriptors.[11] Recent research efforts are directed toward the enhancement of the NMR shift predictions by nonlinear models, such as the neural network model applied to ^{13}C in alkanes and cycloalkanes,[12−15] in monosubstituted benzenes,[16,17] for keto-steroids,[18] and halomethanes[19] as well as for ^{31}P.[20,21]

MultiLayer Feedforward (MLF) Artificial Neural Networks (ANN)[22,23] are a promising model for solving Quantitative Structure−Property Relationships (QSPR) problems, and they are particularly useful in cases where it is difficult to specify an exact mathematical model which describes a specific structure−property relationship. In such cases ANNs, which employ learning procedures based on the patterns describing the molecular structure and the investigated property in order to develop an internal representation of a physicochemical phenomena, may be able to form structural correlations which produce accurate predictions. Also, conventional statistical methods require a specific function over which the data are to be regressed. In order to specify this function, one must derive a mathematical model of the phenomena. Furthermore, considerable mathematical and computational effort is required to obtain convergence if these functions are highly nonlinear.

The recent growing interest in the application of ANNs in the field of QSPR is a result of its demonstrated superiority over the traditional multilinear regression (MLR) or additive models in numerous cases of interest.[24−27] The advantage of ANNs is the presence of hidden layers and the use of nonlinear activation functions which allows nonlinear mapping of the structural parameters to the corresponding physicochemical property. The increased number of adjustable parameters also helps to improve the calibration and prediction accuracy, much like adding another descriptor to a linear model will improve it.

In the present paper we present a new 12-digit code for the environment of the sp^2 carbon atom in acyclic alkenes, which has been investigated as a structural descriptor for the prediction of the ^{13}C NMR chemical shift. Two models have been investigated, namely the multilinear regression and the multilayer feedforward artificial neural networks. In order to evaluate the modeling and prediction capability of the two approaches, the calibration and leave-20%-out cross-validation results are compared for both models. Three activation functions have been used in the neural model: the hyperbolic tangent or a bell-shaped function for the hidden layer and a linear or a hyperbolic tangent function for the output layer. The predictions of the neural model are compared with three additive schemes for the calculation of the sp^2 ^{13}C NMR chemical shifts.

[*] Author to whom inquiries about the paper should be addressed.
[†] LARTIC University of Nice-Sophia Antipolis.
[‡] ITODYS.
[§] University "Politehnica" of Bucharest.
[⊗] Abstract published in *Advance ACS Abstracts,* May 1, 1996.

S0095-2338(95)00131-4 CCC: $12.00

2. STUDIED POPULATION AND STRUCTURAL CODING

The studied population encompasses 130 alkenes that correspond to 244 carbon atoms in different environments. The term "environment" means the ranks A, B, C, and D of substituents around the double bond at a topological distance 1−4. Table 1 gathers the data for 25 monosubstituted, 76 disubstituted (30 gem, 23 E, and 23 Z), 22 trisubstituted (9 gem, 7 Z, and 6 E), and 6 tetrasubstituted alkenes. It is well established that the proximate environment (A,B positions) has the most important effect upon sp^2 ^{13}C chemical shifts. Thus 38 different A,B environments are represented in the studied population.

The topological structural encoding used here already proved its effectiveness by both multilinear relationships[10] and in neuromimetic approaches[13] for ^{13}C chemical shifts prediction. These topological descriptors lie on a strictly ordered description taking into account the connectivity of the sites and their (possible) chromatism. Such a structural encoding only expresses topological (not geometrical) distances. So, in the particular case of alkene description, additional information is needed to specify *Z*, *E* stereoisomerism. We therefore tested a mode of description allowing for directly expressing the relative topographical location of the sites around the double bond. We need two sets of rules: ordering rules and description rules.

Ordering rules: For A positions (around the resonating carbon), sites are hierarchically ordered according to

−whether site A is occupied by a carbon atom,

−greater number of B neighbors on A sites,

−location of the considered branch (*E*/*Z* with respect to the prioritary A′ position on the other sp^2 carbon atom),

−greater number of occupied C positions,

−greater number of occupied D positions.

For A′ positions (on the alternate sp^2 carbon atom), sites are hierarchically ordered according to

−position E with respect to A1 site,

−greater number of occupied B and successively C and D positions.

Once the environment is ordered, the description rules make possible the expression of the organization of the environment thanks to a set of 12 parameters: A_1 site (0 or 1), number of B_1, A_2 site, number of B_2, A'_1 site, number of B'_1, A'_2 site, number of B'_2, number of C, number of D, number of C′, and number of D′. This encoding is not biunequivocal but seems justified by the limited influence of C and D positions on the chemical shifts. This decription spans over sites up to a topological distance of 5 across the double bond, although the influence of these positions on the chemical shifts remains quite small. We choose this encoding to maintain a symmetrical description of the environments for the two sp^2 carbon atoms. The site labeling is given in Figure 1 by letters and indices. All site categories in the environment of the resonating carbon atom are well represented in the studied population, as evidenced by their occurrence shown in Figure 1 by italicized numbers.

Example of descriptors for the common substituents are given in Table 2. The IUPAC name, substituents, and descriptors of the complete studied population are given in Excel files offered as supporting information (online only).

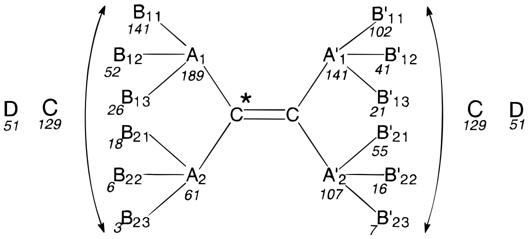


Figure 1. Number of occupied sites in the studied population.

3. METHOD

Network Architecture. We have used multilayer feed-forward neural networks provided with a single hidden layer of neurons. While the size of the input layer of the network is determined by the length of the code used to describe the environment of the resonating carbon (i.e., 12 in the present study), the number of neurons in the hidden layer was selected on the basis of systematic empirical trials in which ANNs with increasing number of hidden neurons were trained to predict the experimental ^{13}C chemical shift values ($\delta_{\rm exp}$). We have used a bias neuron connected to all neurons in the hidden and output layers, and one output neuron which provides the calculated value of the ^{13}C chemical shift.

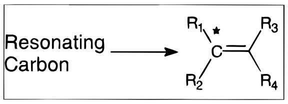
Activation Functions. The usual activation function (logistic function) has a sigmoidal shape (Figure 2a) and takes values between 0 and 1; for large negative arguments its value is close to 0, and the practice demonstrated that learning is difficult in such conditions. To overcome this deficiency of the logistic function, in the present study we have used the hyperbolic tangent (tanh) which takes values between −1 and 1 (Figure 2b). For the hidden layer neurons we have also investigated a bell-shaped activation function (Figure 2c), defined as $\mathrm{Act}(z) = 1/(1+z^2)$. The bell function offered better ANN models than a sigmoidal function in QSPR studies where a highly nonlinear relationship exists between structural parameters and the investigated property.[14,28,29] In certain situations, a linear output function (Figure 2d) provides better quantitative estimations than a sigmoidal one; therefore, for the output layer we have investigated both the linear and the tanh activation functions.

Software Used. The program was written in Borland C language and run on a PC486 DX2 at 66MHz. Typical time of execution for a complete training is about 10−20 min.

Data Set. The structure and experimental ^{13}C chemical shift of sp^2 atoms in 130 acyclic alkenes used in the present investigation were taken from the literature and are presented in Table 1 columns 2−6. A total of 244 structurally unique sp^2 carbon atoms from these compounds were used to constitute the patterns of the data set, where a pattern consists in a pair *p(Iv,T)* made of an input vector *Iv* with 12 components (encoding the environment of the resonating carbon) and a target value *T* (which is the experimental ^{13}C chemical shift $T = \delta_{\rm exp}$). Each component of the input vector and the target values were scaled between −0.9 and 0.9. The minimum and maximum values of $\delta_{\rm exp}$ used in the calibration of the model were 104.6 and 165.0 ppm, respectively.

Learning Method. The training of the ANNs was performed with the standard backpropagation method,[30] until convergence was obtained, i.e., the correlation coefficient between experimental and estimated values improved by less than 10^{-5} in 100 epochs. The patterns were presented randomly to the network, and the weight updates were made

Table 1. Structure, Experimental ^{13}C Chemical Shift, Mean Calibration and Prediction ANN Residuals, Calibration and Prediction MLR Residuals for the Investigated sp^2 Carbon Atoms



							ANN residuals		MLR residuals	
N	R_1	R_2	R_3	R_4	expe.	ref[a]	calib	pred	calib	pred
1	H	H	H	H	123.5	K	0.05	−0.65	0.41	0.69
2	H	H	Me	H	115.9	K	0.02	−0.16	0.14	0.48
3	Me	H	H	H	133.4	K	−0.04	0.34	0.60	0.87
4	H	H	Me	Me	111.3	K	0.57	0.20	1.78	1.65
5	Me	Me	H	H	141.8	K	−0.08	0.46	2.61	2.43
6	Me	H	H	Me	125.4	K	−0.63	−0.50	−1.16	−1.40
7	Me	H	Me	H	124.2	K	−0.31	−0.69	−1.27	−1.13
8	Me	Me	Me	H	131.6	K	−0.53	0.28	−0.26	−0.34
9	Me	H	Me	Me	118.6	K	−0.53	−0.90	−0.63	−0.96
10	Me	Me	Me	Me	123.5	K	−0.88	−0.75	−2.12	−2.20
11	H	H	Et	H	113.5	K	0.13	0.12	−1.42	−1.41
12	Et	H	H	H	140.5	K	−0.56	−0.38	2.38	2.46
13	H	H	Et	Me	109.1	K	−0.29	−0.47	0.42	0.33
14	Et	Me	H	H	147.0	K	−0.15	−0.38	2.49	2.88
15	Me	H	H	Et	123.5	K	−0.25	0.34	−2.15	−2.34
16	Et	H	H	Me	133.2	K	−0.34	0.28	1.32	1.24
17	Me	H	Et	H	122.8	K	−0.39	−0.76	−1.83	−1.97
18	Et	H	Me	H	132.4	K	0.07	0.23	1.61	1.59
19	Me	H	Me	Et	117.1	C	−0.51	−0.38	−1.22	−1.38
20	Et	Me	H	Me	137.7	C	0.50	0.50	−0.57	−0.58
21	Me	H	Et	Me	118.2	C	−0.36	0.09	−0.19	−0.20
22	Et	Me	Me	H	137.9	C	0.47	0.63	0.72	0.78
23	Et	H	Me	Me	126.9	C	0.61	0.59	2.36	2.34
24	Me	Me	Et	H	130.6	C	−0.25	−0.04	−0.43	−0.45
25	Et	Me	Me	Me	129.6	C	0.15	−0.12	−1.34	−1.34
26	Me	Me	Et	Me	123.1	C	−0.50	−0.86	−1.68	−1.78
27	H	H	iPr	H	111.4	C	0.11	−0.27	−2.69	−2.81
28	iPr	H	H	H	145.9	C	−0.14	0.13	2.46	2.44
29	H	H	iPr	Me	107.9	C	−0.50	−0.20	0.06	0.01
30	iPr	Me	H	H	151.7	C	0.88	1.49	1.87	1.85
31	H	H	Et	Et	106.7	K	0.17	0.04	−1.07	−1.19
32	Et	Et	H	H	152.6	K	−0.56	−0.41	4.46	4.68
33	Me	H	iPr	H	121.9	F	−0.21	−0.14	−1.89	−1.96
34	iPr	H	Me	H	138.9	F	1.11	0.90	2.80	2.53
35	Et	H	H	Et	131.2	K	−0.29	0.11	0.23	0.40
36	Et	H	Et	H	131.2	K	0.08	0.50	1.25	1.15
37	Me	H	Me	iPr	116.2	K	−0.08	0.09	−1.21	−1.32
38	iPr	Me	H	Me	141.5	K	0.18	0.32	−2.09	−2.00
39	Me	H	iPr	Me	117.9	K	−0.38	−0.27	0.35	0.32
40	iPr	Me	Me	H	141.0	K	−0.32	−0.63	−1.50	−1.46
41	Me	H	Et	Et	116.9	F	−0.31	−0.25	−0.58	−0.40
42	Et	Et	Me	H	143.9	F	0.37	0.07	3.09	3.01
43	Et	H	Me	Et	125.2	C	0.18	1.22	1.57	1.56
44	Et	Me	H	Et	136.3	C	−0.31	−0.15	−1.06	−1.37
45	Et	H	Et	Me	126.5	C	0.65	1.58	2.79	2.86
46	Et	Me	Et	H	136.5	C	0.04	0.14	0.16	0.21
47	iPr	H	Me	Me	133.0	C	0.88	1.72	3.14	3.33
48	Me	Me	iPr	H	128.9	C	−0.94	−1.21	−1.29	−1.27
49	Me	Me	Et	Et	123.1	C	0.18	1.46	−0.77	−0.52
50	Et	Et	Me	Me	135.9	C	0.66	2.35	1.34	1.43
51	H	H	tBu	H	108.5	K	−1.19	−1.72	−4.75	−4.92
52	tBu	H	H	H	149.3	K	0.10	0.66	0.55	0.59
53	H	H	tBu	Me	108.3	K	0.58	0.99	1.29	1.30
54	tBu	Me	H	H	153.4	K	0.00	0.46	−1.75	−1.97
55	H	H	iPr	Et	105.4	C	−0.56	−0.40	−1.53	−1.60
56	iPr	Et	H	H	157.6	C	0.52	1.02	4.15	4.02
57	Me	H	H	tBu	119.3	H	0.10	−0.47	−4.53	−4.77
58	tBu	H	H	Me	143.0	H	0.73	1.39	0.49	0.51
59	Me	H	tBu	H	122.5	H	1.24	1.22	−0.46	−0.37
60	tBu	H	Me	H	141.0	H	−0.26	−0.26	−0.42	−0.80
61	Et	H	H	iPr	129.0	C	−0.19	−0.03	−1.06	−0.72
62	iPr	H	H	iPr	136.8	C	−0.40	−0.86	0.52	0.48
63	Et	H	iPr	H	129.3	C	−0.81	−0.64	0.19	0.12
64	iPr	H	Et	H	137.1	C	0.40	0.39	1.83	1.88
65	Me	H	Et	iPr	116.0	H	0.08	0.11	−0.57	−0.83
66	iPr	Et	H	Me	148.1	H	−0.78	−0.77	0.89	1.21
67	Me	H	iPr	Et	116.0	H	−1.03	−0.67	−0.64	−0.64

Table 1 (Continued)

							ANN residuals		MLR residuals	
N	R_1	R_2	R_3	R_4	expe.	ref[a]	calib	pred	calib	pred
68	iPr	Et	Me	H	146.6	H	−1.19	−2.09	0.48	0.49
69	Me	Me	tBu	H	130.0	C	0.90	0.85	0.65	0.74
70	tBu	H	Me	Me	135.2	C	−0.79	0.44	0.02	0.21
71	Me	H	Me	tBu	114.8	C	−0.07	−0.31	−1.70	−1.97
72	tBu	Me	H	Me	143.9	C	−0.47	−0.01	−5.01	−5.17
73	H	H	iPr	iPr	104.6	D	−2.51	−4.21	0.32	1.04
74	iPr	iPr	H	H	162.3	D	−0.87	−0.92	−3.72	−3.92
75	iPr	H	H	iPr	134.9	K	−0.48	−1.45	−0.47	−0.54
76	iPr	H	iPr	H	135.4	K	−0.43	−0.44	0.97	0.72
77	tBu	Et	H	H	158.9	D	−0.97	−0.43	0.13	0.19
78	Et	H	H	tBu	126.5	C	−0.12	−0.40	−2.65	−2.34
79	tBu	H	H	Et	140.6	C	−0.28	−1.51	−1.00	−0.96
80	Et	H	tBu	H	130.7	C	1.41	1.74	2.42	2.45
81	tBu	H	Et	H	139.2	C	−1.03	−1.43	−1.39	−1.16
82	H	H	tBu	iPr	104.9	D	−0.58	−0.04	−0.29	−0.34
83	tBu	iPr	H	H	165.0	D	1.36	2.57	2.60	2.49
84	H	H	tBu	tBu	109.2	D	2.09	2.46	4.92	4.85
85	tBu	tBu	H	H	164.3	D	1.13	1.69	−1.72	−1.65
86	tBu	Me	Me	tBu	137.3	K	−0.29	0.37	−1.54	−2.31
87	H	H	nPr	H	114.5	C	0.30	0.22	−1.52	−1.64
88	nPr	H	H	H	139.0	C	−0.53	−0.13	1.48	1.67
89	H	H	nPr	Me	110.0	D	−0.37	−0.50	0.22	0.49
90	nPr	Me	H	H	145.8	D	−0.33	0.01	1.89	1.94
91	Me	H	H	nPr	124.9	K	0.13	0.19	−1.85	−1.94
92	nPr	H	H	Me	131.7	K	−0.42	−0.07	0.42	0.49
93	Me	H	nPr	H	124.0	K	0.40	0.05	−1.73	−1.86
94	nPr	H	Me	H	130.8	K	−0.06	−0.21	0.61	0.77
95	Me	H	Me	nPr	118.4	K	0.20	−0.16	−1.02	−1.09
96	nPr	Me	H	Me	135.9	K	−0.29	−0.45	−1.77	−1.83
97	Me	H	nPr	Me	119.2	K	0.08	−0.26	−0.29	−0.45
98	nPr	Me	Me	H	136.1	K	−0.38	−0.06	−0.48	−0.43
99	Me	Me	nPr	H	131.3	K	1.82	2.68	0.87	1.05
100	nPr	H	Me	Me	125.0	K	−1.67	−2.53	−0.65	−0.90
101	Me	Me	nPr	Me	123.9	K	0.10	0.22	−1.99	−1.71
102	nPr	Me	Me	Me	128.0	K	−0.55	−1.28	−2.34	−2.31
103	H	H	sBu	H	112.5	K	0.32	0.37	−2.69	−2.57
104	sBu	H	H	H	144.7	K	−0.25	0.12	1.86	1.91
105	H	H	sBu	Me	109.5	K	0.07	−0.03	0.56	0.55
106	sBu	Me	H	H	150.0	K	−0.02	0.10	0.77	0.92
107	H	H	nPr	Et	107.8	K	0.27	−0.31	−1.07	−1.22
108	nPr	Et	H	H	151.5	K	−0.63	−0.74	3.96	3.86
109	Me	H	H	sBu	123.0	K	0.36	0.28	−2.84	−2.91
110	sBu	H	H	Me	137.6	K	−0.27	0.13	1.01	1.30
111	Me	H	sBu	H	122.5	K	−0.09	−0.59	−2.40	−2.59
112	sBu	H	Me	H	137.3	K	0.53	0.49	1.80	1.94
113	Et	H	H	nPr	132.4	H	−0.31	0.56	0.33	0.33
114	nPr	H	H	Et	129.4	H	−0.53	−0.42	−0.97	−0.74
115	Et	H	nPr	H	131.9	H	0.14	0.11	0.85	1.00
116	nPr	H	Et	H	129.2	H	−0.42	−0.63	−0.15	−0.06
117	Me	H	Me	sBu	118.1	H	1.12	1.01	−0.41	−0.48
118	sBu	Me	H	Me	139.9	H	−0.64	−0.63	−3.09	−3.29
119	Me	H	sBu	Me	119.4	H	0.37	0.30	0.75	0.82
120	sBu	Me	Me	H	139.6	H	−1.00	−0.74	−2.30	−2.53
121	H	H	tPent	H	110.3	C	−0.36	−0.80	−4.05	−4.23
122	tPent	H	H	H	148.4	C	0.03	0.56	0.25	0.28
123	H	H	tPent	Me	109.6	C	0.76	0.75	1.49	1.57
124	tPent	Me	H	H	152.1	C	−0.64	−0.34	−2.45	−2.48
125	H	H	sBu	Et	107.2	D	0.16	−0.29	−0.83	−1.06
126	sBu	Et	H	H	155.3	D	−1.02	−0.95	2.45	2.50
127	H	H	tpent	tBu	111.2	D	2.05	0.48	5.82	6.44
128	tPent	tBu	H	H	161.4	D	−1.54	−1.75	−4.02	−4.27
129	H	H	iBu	H	115.4	C	0.21	−0.10	−1.72	−2.08
130	iBu	H	H	H	137.8	C	−0.09	0.25	0.88	1.07
131	H	H	iBu	Me	111.3	C	−0.28	−0.51	0.42	0.66
132	iBu	Me	H	H	144.8	C	−0.25	0.13	1.49	1.33
133	Me	H	H	iBu	125.8	C	−0.35	−0.45	−2.05	−2.22
134	iBu	H	H	Me	130.4	C	−0.23	0.09	−0.27	−0.87
135	Me	H	iBu	H	124.4	C	0.13	−0.88	−2.43	−2.66
136	iBu	H	Me	H	129.7	C	0.41	0.18	0.12	−0.43
137	Me	Me	iBu	H	131.7	C	−0.36	−0.61	−1.53	−1.89
138	iBu	H	Me	Me	123.9	C	0.34	0.14	0.56	0.71
139	H	H	CH(Me,iPr)	H	113.3	C	0.02	−0.33	−2.99	−3.30
140	CH(Me,iPr)	H	H	H	143.2	C	−0.59	−0.39	0.97	1.11

Table 1 (Continued)

							ANN residuals		MLR residuals	
N	R$_1$	R$_2$	R$_3$	R$_4$	expe.	ref[a]	calib	pred	calib	pred
141	H	H	CH(Et,Et)	H	114.4	C	1.12	1.23	−1.89	−1.84
142	CH(Et,Et)	H	H	H	143.1	C	−0.69	−0.36	0.87	0.96
143	H	H	CH(Et,Et)	Me	111.6	C	0.88	0.77	1.56	1.44
144	CH(Et,Et)	Me	H	H	147.4	C	−1.79	−1.46	−1.23	−1.16
145	nPr	H	H	nPr	130.6	H	−0.47	0.23	−0.87	−0.89
146	nPr	H	nPr	H	130.1	H	−0.09	−0.06	−0.35	−0.55
147	H	H	C(Me,Me,iPr)	Me	110.0	D	−0.22	−0.60	0.79	0.60
148	C(Me,Me,iPr)	Me	H	H	152.0	D	−0.05	0.34	−1.95	−2.17
149	H	H	neoPent	H	116.5	K	0.13	0.02	−1.72	−1.69
150	neoPent	H	H	H	136.1	K	−0.05	0.25	−0.22	−0.72
151	H	H	neoPent	Me	113.8	C	0.79	0.41	1.82	2.03
152	neoPent	Me	H	H	143.9	C	0.00	0.61	1.19	1.23
153	H	H	C(Me,Me,tBu)	Me	112.9	D	1.03	1.00	2.59	3.11
154	C(Me,Me,tBu)	Me	H	H	150.9	D	−0.44	0.43	−2.45	−2.55
155	H	H	C(Me,Me,tBu)	iPr	110.5	D	−0.73	−1.46	2.01	2.20
156	C(Me,Me,tBu)	iPr	H	H	162.7	D	0.61	1.30	2.11	2.30
157	H	H	C(Me,Me,tBu)	tBu	115.7	D	1.02	1.84	8.12	9.16
158	C(Me,Me,tBu)	tBu	H	H	161.7	D	−0.63	−0.25	−2.52	−3.03
159	H	H	neoPent	neoPent	115.8	D	−0.29	−0.07	1.43	1.54
160	neoPent	neoPent	H	H	145.3	D	−0.35	0.30	0.77	1.36
161	neoPent	neoPent	neoPent	neoPent	136.6	K	0.01	0.62	0.79	1.22
162	H	H	nBu	H	114.2	K	0.10	0.28	−0.42	−0.50
163	nBu	H	H	H	139.2	K	−0.47	−0.25	1.22	1.00
164	Me	H	H	nBu	124.7	C	0.18	0.47	−0.65	−0.70
165	nBu	H	H	Me	131.8	C	−0.49	−0.18	0.06	−0.08
166	Me	H	nBu	H	123.7	C	0.07	−0.32	−0.63	−0.76
167	nBu	H	Me	H	131.0	C	0.00	−0.18	0.35	0.42
168	H	H	nBu	Me	109.8	C	−0.21	−0.27	1.42	1.46
169	nBu	Me	H	H	146.1	C	−0.34	−0.20	1.72	2.34
170	Et	H	H	nBu	132.3	H	−0.02	0.44	1.64	1.73
171	nBu	H	H	Et	129.7	H	−0.20	0.06	−1.13	−1.03
172	Et	H	nBu	H	131.7	H	−0.01	0.19	2.05	2.15
173	nBu	H	Et	H	129.5	H	−0.22	0.03	−0.31	−0.34
174	Me	Me	nBu	H	131.1	C	−0.19	0.02	0.38	0.81
175	nBu	H	Me	Me	125.2	C	0.09	0.48	0.79	0.93
176	H	H	sPent	Me	109.3	C	0.32	0.47	1.76	2.03
177	sPent	Me	H	H	150.2	C	−0.13	−0.13	0.51	1.01
178	H	H	nBu	Et	108.0	Sa	0.90	0.21	0.53	0.49
179	nBu	Et	H	H	152.0	Sa	−0.37	−0.43	4.00	4.59
180	H	H	sPent	H	112.3	C	0.38	0.17	−1.48	−1.61
181	sPent	H	H	H	145.1	C	−0.03	0.25	1.80	1.67
182	H	H	CH$_2$CH(Me,Et)	H	115.4	C	0.34	0.06	−0.32	−0.05
183	CH$_2$CH(Me,Et)	H	H	H	137.8	C	−0.22	0.19	0.42	0.53
184	nPr	H	H	nBu	130.5	C	−0.24	−0.11	0.44	0.27
185	nBu	H	H	nPr	130.9	C	−0.14	0.31	−1.03	−0.65
186	nPr	H	nBu	H	129.9	C	−0.27	−0.05	0.85	0.97
187	nBu	H	nPr	H	130.3	C	0.00	0.15	−0.61	−0.87
188	H	H	CH$_2$C(Me,Me,Et)	H	116.5	C	0.31	0.21	−0.32	−0.67
189	CH$_2$C(Me,Me,Et)	H	H	H	135.9	C	−0.36	0.12	−0.88	−0.96
190	H	H	iPent	H	114.1	C	0.06	0.15	0.88	1.28
191	iPent	H	H	H	139.3	C	−0.49	−0.04	0.85	1.59
192	nBu	H	H	nBu	130.5	C	−0.23	0.02	−0.03	−0.14
193	nBu	H	nBu	H	130.2	H	−0.09	−0.18	0.69	0.93
194	H	H	nPent	H	114.2	C	0.10	0.15	−0.42	−0.11
195	nPent	H	H	H	139.2	C	−0.47	−0.02	1.22	1.13
196	Me	H	H	nPent	124.7	H	0.18	0.45	−0.65	−0.84
197	nPent	H	H	Me	132.0	H	−0.29	0.17	0.26	0.57
198	Me	H	nPent	H	123.7	H	0.07	0.12	−0.63	−0.64
199	nPent	H	Me	H	130.7	H	−0.30	−0.30	0.05	0.13
200	Et	H	H	nPent	132.0	C	−0.32	−0.24	1.34	1.36
201	nPent	H	Et	H	129.5	C	−0.40	0.00	−1.33	−1.14
202	Et	H	nPent	H	131.8	H	0.09	0.29	2.15	2.25
203	nPent	H	Et	H	129.6	H	−0.12	0.29	−0.21	−0.37
204	nPr	H	H	nPent	130.3	C	−0.44	−0.31	0.24	0.07
205	nPent	H	H	nPr	130.8	C	−0.24	0.19	−1.13	−1.13
206	nPr	H	nPent	H	129.9	C	−0.27	−0.05	0.85	0.97
207	nPent	H	nPr	H	130.5	C	0.20	0.35	−0.41	−0.67
208	Me	Me	C$_2$H$_4$CH(Me,Et)	H	125.3	C	0.05	0.31	0.43	0.47
209	C$_2$H$_4$CH(Me,Et)	H	Me	Me	130.8	C	−0.57	−0.31	1.48	1.78
210	H	H	nHex	Me	109.8	C	−0.21	−0.24	1.42	1.21
211	nHex	Me	H	H	146.1	C	−0.34	−0.20	1.72	2.34
212	Me	H	H	nHex	124.7	H	0.18	0.47	−0.65	−0.70
213	nHex	H	H	Me	132.0	H	−0.29	0.02	0.26	0.12

Table 1 (Continued)

							ANN residuals		MLR residuals	
N	R_1	R_2	R_3	R_4	expe.	ref[a]	calib	pred	calib	pred
214	Me	H	nHex	H	123.7	H	0.07	−0.03	−0.63	−0.53
215	nHex	H	Me	H	131.2	H	0.20	0.39	0.55	0.47
216	Et	H	H	nHex	132.4	H	0.08	0.54	1.74	1.83
217	nHex	H	H	Et	129.9	H	0.00	0.26	−0.93	−1.11
218	Et	H	nHex	H	131.9	H	0.19	0.52	2.25	2.32
219	nHex	H	Et	H	129.7	H	−0.02	0.23	−0.11	−0.14
220	nHex	H	H	nHex	130.0	Sa	−0.73	−0.48	−0.53	−0.64
221	Me	H	H	nHept	124.6	C	0.08	0.37	−0.75	−0.80
222	nHept	H	H	Me	131.8	C	−0.49	−0.03	0.06	0.37
223	Me	H	nHept	H	123.8	H	0.17	−0.08	−0.53	−0.40
224	nHept	H	Me	H	131.2	H	0.20	0.39	0.55	0.47
225	H	H	nHept	Me	109.8	C	−0.21	−0.21	1.42	1.79
226	nHept	Me	H	H	146.0	C	−0.44	−0.32	1.62	1.56
227	H	H	nHex	H	114.2	C	0.10	0.18	−0.42	−0.56
228	nHex	H	H	H	139.2	C	−0.47	0.17	1.22	1.13
229	H	H	nNon	Me	110.0	Sa	−0.01	−0.01	1.62	1.99
230	nNon	Me	H	H	146.0	Sa	−0.44	−0.30	1.62	2.24
231	H	H	C(Bu,Et,Me)	H	111.6	Sc	0.26	0.21	−2.45	−2.37
232	C(Bu,Et,Me)	H	H	H	146.8	Sc	−0.90	−0.78	−1.22	−1.28
233	H	H	C(Bu,Et,Et)	H	112.4	Sc	−0.33	−0.43	−2.75	−3.24
234	C(Bu,Et,Et)	H	H	H	146.6	Sc	−0.19	0.00	−0.82	−0.85
235	H	H	C(Bu,Me,Pr)	H	111.2	Sc	0.28	−0.07	−1.44	−1.86
236	C(Bu,Me,Pr)	H	H	H	147.3	Sc	−0.59	−0.53	−1.18	−0.87
237	H	H	C(Bu,Et,Pr)	H	112.2	Sc	−0.02	−0.06	−1.54	−1.50
238	C(Bu,Et,Pr)	H	H	H	146.8	Sc	−0.16	0.26	−1.08	−0.97
239	H	H	C(Bu,Bu,Me)	H	111.2	Sc	0.28	−0.01	−1.44	−1.52
240	C(Bu,Bu,Me)	H	H	H	147.3	Sc	−0.59	−0.51	−1.18	−1.33
241	H	H	C(Bu,Pr,Pr)	H	111.9	Sc	0.10	0.06	−0.44	−0.44
242	C(Bu,Pr,Pr)	H	H	H	147.1	Sc	−0.03	0.41	−1.24	−1.26
243	H	H	C(Bu,Bu,Et)	H	112.2	Sc	−0.02	−0.32	−1.54	−2.07
244	C(Bu,Bu,Et)	H	H	H	146.9	Sc	−0.06	0.30	−0.98	−0.87

[a] C: Couperus, P. A.; Clague, A. D. H.; van Dongen, J. P. C. M. ^{13}C Chemical Shifts of some Model Olefins. *Org. Magn. Reson.* **1976**, *8*, 246−431. D: Dubois, J.-E.; Carabedian, M. Modeling of the Alkyl Environment Effects on ^{13}C Chemical Shifts. *Org. Magn. Reson.* **1980**, *5*, 264−271. F: Friedel, R. A.; Retcofsky, H. L. Carbon-13 Nuclear Magnetic Resonance Spectra of Olefins and Other Hydrocarbons. *J. Am. Chem. Soc.* **1963**, *85*, 1300−1306. H: de Haan, J. W.; van de Ven, L. J. M. Configurations and Conformations in Acyclic, Unsaturated Hydrocarbons. A 13C NMR Study. *Org. Magn. Reson.* **1973**, *5*, 147−153. K: Kalinowski, H.-O.; Berger, S.; Braun, S. *Carbon-13 NMR Spectroscopy*; John Wiley & Sons: Chichester, 1988; pp 132−134. Sa: ^{13}C NMR Database from Bio-Rad Laboratories, Inc. Sadtler Division. Philadelphia, PA 19104, U.S.A. Sc: Schwartz, R. M.; Rabjohn, N. ^{13}C NMR Chemical Shifts of Some highly-branched Acyclic Compounds. *Org. Magn. Reson.* **1980**, *13*, 9−13.

Table 2. Descriptors for Substituents up to Five Carbons

	A	B	C	D	note
H	0	0	0	0	
Me	1	0	0	0	
Et	1	1	0	0	
nPr	1	1	1	0	
iPr	1	2	0	0	
nBu	1	1	1	1	*a*
iBu	1	1	2	0	
sBu	1	2	1	0	
tBu	1	3	0	0	
nPent	1	1	1	1	*a*
iPent	1	1	1	2	
$CH_2CH(Me,Et)$	1	1	2	1	
neoPent	1	1	3	0	
sPent	1	2	1	1	
$CHEt_2$	1	2	2	0	*b*
CH(Me,iPr)	1	2	2	0	*b*
tPent	1	3	1	0	

[a] Degeneracy of the code originating from the limitation to a topological distance of 4. [b] Degeneracy of the code originating from the summation over B and C locations.

on a per pattern base. For the initial weights we have used random values between −0.1 and 0.1. In order to evaluate the effect of the initial weights seed 10 different sets have been generated and trained to convergence for each network investigated. The momentum was set to 0.8, and the learning rate was maintained constant during training but depends on the activation function used.

Performance Evaluation. Two types of performances have to be evaluated: model calibration and prediction. The quality of model calibration is estimated by comparing the chemical shift values calculated (δ_{calc}) during the training phase and target values (δ_{exp}), while the predictive power of the network (architecture + activation functions) is estimated by a cross-validation method.[31]

Performance Indicators. In order to compare the performance of the ANN models with the statistical results of the MLR equation, we have used the correlation coefficient *r*, the standard deviation *s*, and the mean residual mres of the linear correlation between experimental δ_{exp} and calculated δ_{calc} chemical shifts: $\delta_{exp} = A + B\cdot\delta_{calc}$. The mean residual was calculated by the formula

$$\text{mres} = \sum_i \frac{|(\delta_{exp} - \delta_{calc})_i|}{n}$$

where *n* represents the number of patterns. We determined for each model the number and residuals of statistical outliers, i.e., those cases with an absolute difference between δ_{exp} and δ_{calc} greater than three times the standard deviation.

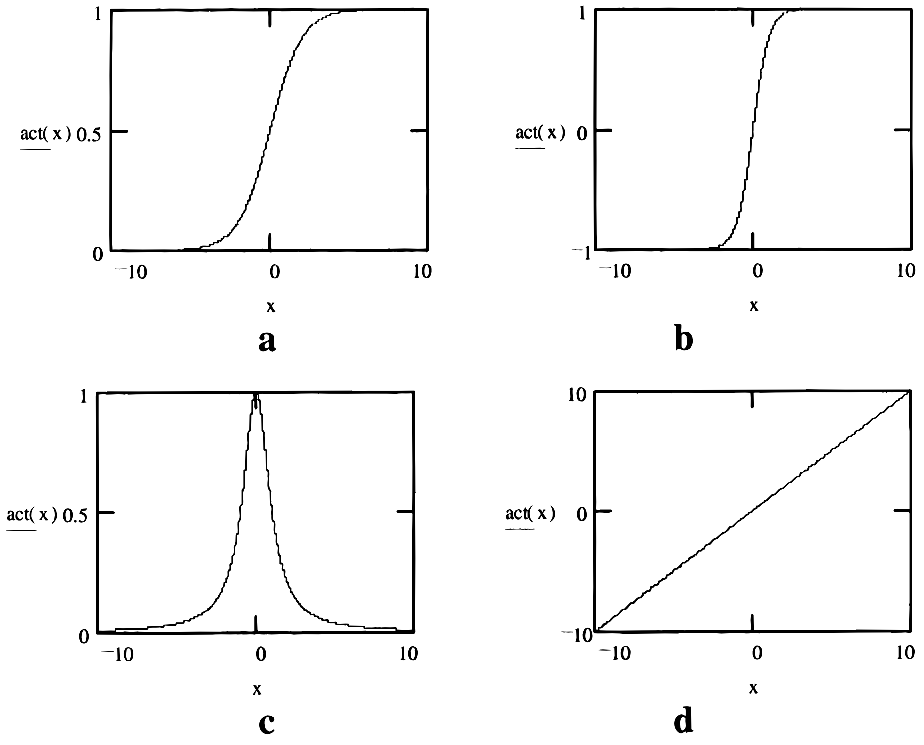


Figure 2. Activation functions: a, logistic; b, hyperbolic tangent; c, bell; and d, linear.

Cross-Validation. The goal of a QSPR study is to develop a model of the investigated phenomena that can give reliable predictions for new patterns that were not used in the calibration of the mathematical model. In the present study we have used a leave-20%-out (L20%O) cross-validation method in order to estimate the predictive capabilities of the models. The L20%O cross-validation was applied by forming a prediction set of patterns which consisted of 20% of the patterns selected at random from the entire set of 244 data, and then the ANN model was calibrated with a learning set consisting of the remaining 80% of the data. Finally, the neural model obtained in the calibration phase was used to predict the chemical shift value for the patterns in the prediction set. This procedure is repeated five times, until all patterns are selected in a prediction set once and only once. A linear regression between experimental δ_{exp} and predicted δ_{calc} allows one to obtain the statistical indices used to compare the prediction capabilities of different ANN architectures and selections of activation functions. The complete procedure was repeated several times using a different partioning to ensure that consistent results were obtained.

4. RESULTS AND DISCUSSION

ANN Architectures. We designed a set of experiments to study the influence of four different factors on the performance of neural networks trained with the back propagation algorithm: the number of hidden neurons, the activation functions of the hidden and output layers, the initial weights, and the learning rates.

Because there is no theoretical way to establish the optimal number of hidden neurons and the type of activation function for a MLF ANN, we have investigated a large number of networks with different characteristics. The number of hidden neurons was varied between 2 and 7, and we have used the following combinations of hidden activation function, output activation functions, and their respective learning rates: (tanh, tanh, 0.01, 0.01), (tanh, linear, 0.01, 0.01), (bell, tanh, 0.01, 0.01), and (bell, linear, 0.05, 0.01). Excellent results were obtained in all cases investigated, and even for networks with only two hidden neurons the correlation coefficient ***r*** is higher than 0.996 and the standard deviation ***s*** is lower than 1.2 ppm. For a four-neurons hidden layer ***r*** equals 0.999, and ***s*** ranges from 0.59 to 0.63 ppm. The statistical quality of the model did not improve significantly for the networks with up to seven hidden neurons.

The modeling power of the ANN model can be improved by using networks with a higher number of hidden neurons, but in QSPR studies it is important to take into consideration that MLF networks are universal approximators,[32] which are capable of arbitrarily accurate approximation to arbitrary mappings, provided a sufficiently number of hidden units is available. Recently, Andrea and Kalayeh[33] investigated the potential of chance correlations in QSPR models and proposed to characterize a network by a structural parameter ρ, which is the ratio of the number of patterns in the training set to the number of connections. Based on empirical observations, Livingstone[34] proposed to use for QSPR studies only networks with a ρ parameter greater than 2, in order to ensure that the network can give reliable predictions. The network with four hidden neurons has 57 connections, and a ρ value equal to 4.3, which ensures that the chance correlations are unlikely to appear in our case. On the other hand, the standard error of the neural model is comparable with the experimental error of the ^{13}C chemical shift used to develop the model, because we have used data determined in different laboratories. Table 3 presents the statistical indices and outliers of the ANNs calibration models with four hidden neurons, for the four combinations of activation functions investigated. In each case, we report only the optimal network obtained for the 10 different sets of initial weights generated. Although the different trial sets generally

Table 3. Statistical Results, Outliers and Their Residuals for ANNs with Four Neurons in the Hidden Layer and Different Combinations of Activation Functions Obtained for the Calibration Phase

ANN calibration										
hidden layer		output layer		statistical results						
activation function	learning rate	activation function	learning rate	epochs	A	B	*r*	*s*	*mres*	outliers (num,residual)
bell	0.05	linear	0.01	1000	0.534	0.995	0.999	0.598	0.429	(73,−2.2),(84,2.4), (99,2.1),(127,2.4)
bell	0.01	tanh	0.01	1300	0.691	0.994	0.999	0.612	0.460	(73,−3.1),(99,2.0)
tanh	0.01	linear	0.01	1800	1.926	0.985	0.999	0.595	0.436	(73,−2.2),(84,2.4),(99,1.8),(100,−1.8),(127,2.2), (128,−2.0),(144,−1.8)
tanh	0.01	tanh	0.01	3000	1.809	0.985	0.999	0.634	0.504	(73,−2.5),(83,2.4),(84,2.1),(127,2.0),(144,2.0)
mean of the four ANNs					1.197	0.990	0.999	0.555	0.415	(73,−2.5),(84,2.1),(99,1.8),(100,−1.7),(127,2.1),(144,-1.8)

Table 4. Statistical Results, Outliers and Their Residuals for ANNs with Four Neurons in the Hidden Layer and Different Combinations of Activation Functions Obtained for the Prediction by the Leave-20%-Out Cross-Validation Method

ANN prediction 20% out									
hidden layer		output layer		statistical results					
activation function	learning rate	activation function	learning rate	A	B	*r*	*s*	*mres*	outliers (num,residual)
bell	0.05	linear	0.01	−0.482	1.004	0.998	0.919	0.630	(1,5.0),(73,−4.7),(99,3.0)
bell	0.01	tanh	0.01	0.289	0.998	0.997	1.066	0.664	(1,−6.2),(73,−4.7),(161,−6.0)
tanh	0.01	linear	0.01	0.649	0.995	0.998	0.894	0.581	(73,−3.9),(84,3.3),(99,2.8),(157,3.9),(161,3.7)
tanh	0.01	tanh	0.01	−0.492	0.996	0.997	1.067	0.712	(50,4.2),(66,−3.9),(68,−3.3),(73,−3.6),(83,3.4), (157,5.0),(161,3.9)
mean of the four ANN				−0.040	1.000	0.999	0.753	0.499	(50,2.4),(73,−4.2),(83,2.6),(84,2.5),(99,2.7),(100,−2.5)

give close results, it may happen that, for a particular set of initial weights, results of slightly poorer quality are obtained. Thus it seems important to recommend that in such application one should not limit the investigation to a single trial with one set of random initial weights, but make at least two trials which should give close results.

From a practical point of view, a QSPR model is as valuable as its predictions are, and the ANN model is known to give poor predictions for new patterns if the network memorizes the training set without extracting the most important features from data. The L20%O cross-validation was applied for all networks investigated, providing reliable information concerning the predictive power of the ANN model. For all the networks, with a dimension of the hidden layer between 2 and 7 neurons, the cross-validation correlation coefficient ***r*** between δ_{exp} and δ_{pred} lies between 0.990 and 0.997, while the cross-validation standard error ***s*** takes values between 0.89 and 2.00 ppm. The L20%O results for the networks provided with four hidden neurons are presented in Table 4. The correlation coefficients ***r*** obtained for the four combinations of activation functions are very close to each other: 0.997−0.998, and the standard error ***s*** lies in the range 0.89−1.07 ppm. The results obtained in our computations indicates that using four hidden neurons offers a good balance between calibration and prediction performances, while using more hidden units increases the complexity of the model without significantly improving the results. In what follows we will present only the results obtained for ANNs with four hidden neurons.

As is apparent from Tables 3 and 4, the ANN model gives excellent calibration and prediction results. However, there are a number of cases in which the neural model offers a computed value which differs significantly from the experimental chemical shift. Such outliers, with a residual greater than three times the standard deviation, are presented in the last column of Tables 3 and 4, respectively, for each type of network investigated. For the calibration models, there is one common outlier, the carbon 73, which belongs to a highly branched alkene having two isopropyl substituents attached to the other sp^2 carbon atom. A significant number of outliers is formed by highly branched, sterically crowded alkenes on one side of the double bond, such as carbon 84 with two *tert*-butyl groups and carbon 127 with a *tert*-pentyl and a *tert*-butyl groups attached to the other sp^2 carbon. Residuals observed for two gem tertiary carbons (such as carbon 73) are negative while they are positive for two gem quaternary carbons (such as carbon 84 and 127). As a result compounds with one tertiary and one quaternary carbon attached to the same side of the double bond are not identified as outliers. For example, for the carbon 82 with a *tert*-butyl and an isopropyl groups on the other sp^2 carbon atom we obtain the following residuals (bell-linear: −0.5, bell-tanh: −0.5, tanh-linear: −0.4, tanh-tanh: −1.0). These results suggest that in such cases of alkenes with highly branched substituents the topographical description alone is insufficient to account for the relationship betwen the chemical structure and the chemical shift.

The same trend is shown by the prediction results: we find the same common outlier in the predictions of all four networks, namely carbon 73. The other outliers are specific to a given combination of activation functions. The largest residuals are observed in the prediction of the chemical shift for the carbons of ethylene, which is an extreme case since it is the only molecule having no sp^3 carbon atom. In this case, the different ANNs give very different predicted values (bell-linear: +5.0, bell-tanh: −6.2, tanh-linear: +0.2, tanh-tanh: −1.6) demonstrating that each ANN should be used with extreme caution for extrapolation for compounds which are situated at the limit of the topographical model. Several other cases exhibit such large dispersion in the predicted values by individual ANNs, the more noticeable being atoms 161 (tetraneopentylethylene) and 157 (C1 of 2-(1,1-dimethylethyl)-3,3,4,4-tetramethyl-1-pentene).

Table 5. Statistical Results: Regression Coefficients RC_i, Confidence Interval at 95% Confidence Level CI_i, and Partial Correlation Coefficients r_i of MLR Analysis

parameter i	RC_i	CI_i	r_i
base	123.09	±16.87	
A1	9.71	±1.33	0.712
B1	5.32	±0.73	0.830
A2	6.39	±0.88	0.571
B2	3.62	±0.50	0.487
A′1	−7.33	±1.00	−0.658
B′1	−0.84	±0.11	−0.646
A′2	−6.24	±0.85	−0.403
B′2	−0.91	±0.12	−0.275
C	−0.60	±0.08	0.448
D	0.46	±0.06	0.256
C′	1.10	±0.15	−0.437
D′	−1.40	±0.19	−0.304

The statistical results of the calibration and prediction show little preference for any one combination of activation functions, since in all four cases the correlation coefficients and standard errors are very close. Only the pair (bell-tanh) offered slightly poorer prediction results.

Combined Usage of the Four ANNs. The fact that the four networks give different predicted values in some cases leads us to combine these values by calculating their mean. This procedure has been applied both on the calibration and prediction results. As expected, better statistical indices are obtained, as is apparent from the last rows in Tables 3 and 4. Also, for the prediction by cross-validation, there are only six outliers, namely carbons 50, 73, 83, 84, 99, and 100. Only one carbon (73) gives a residual greater than 3.0 ppm. The residuals between the experimental and mean values for calibration and prediction are reported for the 244 carbons investigated in Table 1, columns 8 and 9.

Comparison with MLR. The primary advantage of using ANNs in QSPR is their capability to provide nonlinear mapping of the structural parameters to the corresponding physicochemical property. In order to compare the ANN model with the linear one, we have used the same topographical structural descriptors and experimental data in a multilinear regression and obtained for calibration a correlation coefficient r equal to 0.992 and standard error s of 1.8 ppm, which shows that the neural model is capable of better calibration models ($r = 0.999$, $s = 0.56$ ppm) in the investigated case. Due to the high value of the standard error, there are only two outliers with an absolute residual greater than three times the standard error: carbon 127 (res 5.8) and 157 (res 8.12) which were also identified as “difficult cases” for ANN. But, if we consider the number of cases having an absolute residual greater than 3 ppm we find 16 such cases in MLR calibration compared to only one for the bell-tanh ANN and none for the mean ANN calibration.

The coefficients of the MLR model, the 95% confidence level of the coefficients, and the partial correlation coefficients are presented in Table 5. Inspection of the partial correlation coefficients indicates that, as expected, the most important descriptors are A1, B1, A′1, and B′1. One must note that the highest intercorrelation coefficients are generally small. The highest value (A1/B′1) did not exceed 0.654 in absolute value, which indicates that the MLR model is free from colinearity problems. We did not investigate the MLR model with a combination of descriptors.

Since we were interested in comparing the predictive power of the ANN and linear models, we have performed the L20%O cross-validation with the same random partitioning of the patterns in five sets for calibration and five sets for prediction used in the ANN model. The statistical indices for the predicted δ_{calc} values by the MLR model are reported in Table 6. The same two outliers identified in calibration show very large residuals in prediction: carbon 127 (res = 6.4) and 157 (res = 9.2). If we compare the statistics of MLR cross-validation with the statistical indices of the ANN models from Table 4, it is clear that the neural network model outperforms the MLR model and provides superior mapping of the structural code to the chemical shift of sp^2 carbon atoms.

Table 6. Statistical Results for the Prediction by MLR and Results for Three Additive Models

	A	B	r	s	*mres*	*MinRes*	*MaxRes*
MLR prediction	0.208	0.998	0.991	1.90	1.47	−5.17	+9.16
Brouwer model	19.68	0.851	0.988	2.15	1.97	−15.4	+14.2
Pretsch model	21.95	0.835	0.970	3.33	3.02	−13.9	+22.0
Cheng model	7.62	0.948	0.987	2.25	1.87	−9.1	+12.0

Table 7. Number of Poorly Predicted Cases and Practical Reliability at Different Levels of Accuracy by the Different Methods Investigated

	2 ppm		2.5 ppm		3 ppm	
accuracy level	Nb	rel	Nb	rel	Nb	rel
mean ANN prediction	7	97.1	4	98.4	1	99.6
MLR prediction	59	75.8	30	87.7	20	91.8
Brouwer model	71	70.9	63	74.2	55	77.4
Pretsch model	130	46.7	112	50.0	87	64.3
Cheng model	84	65.6	52	78.7	32	86.9

The correlation coefficient of the ANN model being higher than that of the MLR model, we conclude that there is a nonlinear dependence between the code which describes the environment of the resonating carbon, and the chemical shift and consequently the use of ANN is justified.

Comparison with Additive Models. To compare the performances of the ANN model with the additive models calibrated to estimate the ^{13}C NMR chemical shift of sp^2 carbon atoms in alkenes, we selected three models developed by Stothers,[4] Pretsch,[5−8] and Cheng.[9] These three additive models were used to estimate the chemical shift of the 244 carbons from Table 1, and their correlation with the experimental values was determined. The statistical indices, given in Table 6, allow us to conclude that the predictions of the ANN model are better than the estimations of these three additive models.

Since the additive methods were calibrated on different sets of experimental data, our comparisons are only qualitative. Still, from a practical point of view, it is useful to identify the number of cases for which each method give an error in prediction that exceeds a threshold value. The number of poorly predicted cases and the practical reliability (reliability = number of good predictions/total number of cases) for levels of accuracy ranging from 2 to 3 ppm for each method are reported in Table 7. These figures shows the superiority of ANNs over the other methods.

5. CONCLUSION

We have presented an ANN QSPR model for the estimation of the ^{13}C NMR shift of sp^2 carbon atoms in acyclic alkenes, using as structural descriptor a vector made of 12 components encoding the environment of the resonating carbon. In the L20%O cross-validation test, the ANN model proved to be superior to the MLR model obtained with the same set of data.

Three activation functions were tested in the neural model: the hyperbolic tangent or a bell-shaped function for the hidden layer and a linear or a hyperbolic tangent function for the output layer. All four combinations of activation functions give close results in the calibration of the ANN model, while for the prediction a linear output function performs better than a hyperbolic tangent one, but from a statistical point of view one could not choose one network over the others. For the ANNs with four neurons in the hidden layer, the standard deviation for calibration ranges between 0.59 and 0.63 ppm, while for prediction it lies between 0.89 and 1.07 ppm.

The networks provided with a bell-shaped hidden function or with a linear output function provide minor differences in the calibration of the model, when compared with a network with tanh activation functions in both hidden and output layers, but we consider them to be good alternatives to the usual sigmoidal-shape activation functions. On the other hand, the results of the prediction tests (the L20%O cross-validation) are more dependent on the selected pair of activation functions, but no definitive conclusions can be obtained from the data investigated in the present paper. We propose a parallel use of the four ANNs for the prediction of unknown shifts, because the mean value of the four predictions exhibits a smaller number of outliers with lower residuals and better statistical indices.

In conclusion, the ANN approach gives both a useful and simple mathematical model for the prediction of the chemical shift of sp^2 carbon atoms in acyclic alkenes. Our results add to the growing support for the use of ANNs in QSPR studies. Also, the results are superior to MLR analysis and additive models, when judged in statistical terms and reliability of prediction.

Supporting Information Available: Excel files containing (a) name, substituents, topographic code, and experimental values of studied compounds and (b) results obtained for calibration and prediction by all networks are available as supporting information via the Internet. For more information on access via Internet consult the masthead page of a recent issue of this Journal.

REFERENCES AND NOTES

(1) Kalchhauser, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103−108.

(2) Chen, L.; Robien, W. The CSEARCH-NMR Data Base Approach to Solve Frequent Questions Concerning Substituent Effects on ^{13}C NMR Chemical Shifts. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 217−223.

(3) 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.

(4) Brouwer, H.; Stothers, J. B. ^{13}C Nuclear Magnetic Resonance Studies. XX ^{13}C Shieldings of Several Allyl Alcohols. Geometric Dependence of ^{13}C Shieldings. *Can. J. Chem.* **1972**, *50*, 1361−1370.

(5) Fürst, A.; Pretsch, E. A Computer Program for the Prediction of ^{13}C -NMR Chemical Shifts of Organic Compounds. *Anal. Chim. Acta* **1990**, *229*, 17−25.

(6) Fürst, A.; Pretsch, E.; Robien, W. Comprehensive Parameter Set for the Prediction of the ^{13}C -NMR Chemical Shifts of sp^3-Hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1990**, *233*, 213−222.

(7) 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.

(8) Pretsch, E.; Fürst, A.; Badertscher, M.; Bürgin, R.; Munk, M. E. C13Shift: A Computer Program for the Prediction of ^{13}C NMR Spectra Based on an Open Set of Additivity Rules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 291−295.

(9) Cheng, H. N.; Kasehagen, L. J. Integrated Approach for ^{13}C Nuclear Magnetic Resonance Shift Prediction, Spectral Simulation and Library Search. *Anal. Chim. Acta* **1994**, *285*, 223−235.

(10) Panaye, A.; Doucet, J. P.; Fan, B. T. Topological Approach of C13 NMR Spectral Simulation: Application to Fuzzy Substructures. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 258−265.

(11) Jurs, P. C.; Ball, J. W; Anker, L. S.; Friedman, T. L. Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272−278.

(12) Kvasnicka, V. An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63−76.

(13) Doucet, J. P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. Neural Networks and ^{13}C NMR Shift Prediction. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324.

(14) Panaye, A.; Doucet, J. P.; Fan, B. T.; Feuilleaubois, E.; El Azzouzi, S. R. Artificial Neural Network Simulation of ^{13}C NMR Shifts for Methyl Substituted Cyclohexanes. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 129−135.

(15) 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.

(16) Kvasnicka, V.; Sklenák, Š.; Pospíchal, J. Application of Recurrent Neural Networks in Chemistry. Prediction and Classification of ^{13}C NMR Chemical Shifts in a Series of Monosubstituted Benzenes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747.

(17) Kvasnicka, V.; Sklenák, Š.; Pospíchal, J. Application of Neural Networks with Feedback Connections in Chemistry: Prediction of ^{13}C NMR Chemical Shifts in a Series of Monosubstituted Benzenes. *J. Mol. Struct. (Theochem)* **1992**, *277*, 87−107.

(18) 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.

(19) Miyashita, Y.; Yoshida, H.; Yaegashi, O.; Kimura, T.; Nishiyama, H.; Sasaki, S. Non-Linear Modelling of ^{13}C NMR Chemical Shift Data Using Artificial Neural Networks and Partial Least Squares Method. *J. Mol. Struct. (Theochem)* **1994**, *311*, 241−245.

(20) West, G. M. J. Predicting Phosphorus NMR Shifts Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 577−589.

(21) West, G. M. J. Predicting Phosphorus NMR Shifts Using Neural Networks. 2. Factors Influencing the Accuracy of Predictions. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 21−30.

(22) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing*; MIT Press: Cambridge, MA, U.S.A., 1986.

(23) Wasserman, P. D. *Neural Computing. Theory and Practice*; Van Nostrand Reinhold: New York, 1989.

(24) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH: Weinheim, 1993.

(25) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.

(26) Wythoff, B. J. Backpropagation Neural Networks. A Tutorial. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 115−155.

(27) Smits, J. R. M.; Melssen, W. J.; Buydens, L. M. C.; Kateman, G. Using Artificial Neural Networks for Solving Chemical Problems. Part I. Multi-Layer Feed-Forward Networks. *Chemom. Intell. Lab. Syst.* **1994**, *22*, 165−189.

(28) Ivanciuc, O. Artificial Neural Networks Applications. Part 5. Prediction of Solubility of C_{60} in a Variety of Solvents. *Croat. Chem. Acta* **1996**, in press.

(29) 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. *Croat. Chem. Acta* **1996**, in press.

(30) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.

(31) E. V. Thomas, E. V. A Primer on Multivariate Calibration. *Anal. Chem.* **1994**, *66*, 795A−804A.

(32) Hornik, K.; Stinchcombe, M.; White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* **1989**, *2*, 359−366.

(33) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(34) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

CI950131X