Monatshefte für Chemie 130, 1089–1095 (1999)

Monatshefte für Chemie
Chemical Monthly

Neural Network Prediction of ^{13}C NMR Chemical Shifts of Substituted Benzenes[a]

Jens Meiler[1], **Reinhard Meusinger**[1,*], and **Martin Will**[2]

[1] Institut für Analytische Chemie, Universität Leipzig, D-04103 Leipzig, Germany
[2] BASF AG, ZHV/Molekülspektroskopie, D-67056 Ludwigshafen, Germany

Summary. A multi-layer feedforward neural network was used for the prediction and assignment of ^{13}C NMR chemical shifts of substituted benzenes. The back-propagation neural network was trained by supervised learning with the chemical shift values of about 1000 substituted benzenes from literature. The average uncertainty for the prediction of the ^{13}C chemical shifts is as low as 1.1 ppm. In comparison to common incremental methods, essentially better results were obtained for highly substituted systems with interacting substituents.

Keywords. ^{13}C NMR chemical shifts; Artificial neural network; Substituted benzenes; Aromatics.

Bestimmung der ^{13}C-NMR-chemischen Verschiebungen substituierter Aromaten mit Hilfe eines neuronalen Netzes

Zusammenfassung. Es wird eine Methode für die Berechnung der ^{13}C-NMR-chemischen Verschiebungen von aromatischen Kohlenstoffatomen in substituierten Benzolen vorgestellt. Hierfür kam ein mehrschichtiges neuronales Netz mit Fehlerrückführung zum Einsatz, welches mit den Literaturwerten der chemischen Verschiebungen von über 1000 monosubstituierten Aromaten trainiert wurde. Das neuronale Netz ist in der Lage, die ^{13}C-chemischen Verschiebungen in Aromaten unabhängig von der Anzahl ihrer Substituenten genau vorherzusagen. Die durchschnittlichen Abweichungen zu den experimentellen Werten sind kleiner als 1.1 ppm. Die Methode ist insbesondere für die Berechnung der Verschiebungswerte höhersubstituierter Benzole deutlich besser geeignet als die bekannten Inkrementverfahren, was an mehreren Beispielen gezeigt wird.

Introduction

Several different methods exist for predicting ^{13}C NMR chemical shifts. A number of powerful quantum chemical procedures are available [1]. However, these calculations are rather time consuming and not possible for large molecules anyway. Two other methods are mostly used in practice for the prediction of ^{13}C NMR chemical shifts: searching in databases and calculations employing incremental systems. Several databases exist containing a large amount of chemical structures and the accompanying NMR spectra [2, 3], in same cases available *via*

[a] Dedicated to Prof. Dr. Dr. h.c. *Harry Pfeifer* on the occasion of his 70th birthday
* Corresponding author

internet [4]. They allow the rapid search for chemical shifts for a given structure or their calculation by means of statistical methods. However, the quality of a database depends on the number and on the correctness of its entries. The last point is a problem arising from uncertain experimental values as an input of databases. Moreover, database research requires the availability of large computer-stored systems, and they are sometimes cost expensive and incomplete. On the other hand, the use of substituent induced chemical shifts (SCS) for the prediction of ^{13}C NMR chemical shifts of different classes of compounds is well-known [5]. Aromatic compounds, especially substituted benzenes, have been intensively investigated. Already in 1979, *Ewing* compiled the SCS of more than 700 mono-substituted benzenes [6].

Several computer programs have been developped using additivity rules including the four increments I_{ipso}, I_{ortho}, I_{meta} and I_{para} for the calculation of chemical shifts of aromatic compounds [7–9]. This method is very powerful when dealing with mono-substituted benzenes and with substituents without sterical interference. In the presence of serious steric and electronic substituent interactions, however, large deviations from additivity in the case of *ortho*- and *para*-substitution and of *ortho*-disubstitution have been observed [9]. The additivity rule is limited at this point, and the calculation of the ^{13}C chemical shifts is a serious problem. In consequence, some authors have proposed the use of numerous correction terms [10, 11].

A relatively new method for the treatment of chemical structures und NMR chemical shift values is comprised by the use of neural networks. Unlike traditional techniques, the aim of neural networks is not to analyze and understand data, but to use them to predict and classify. Artificial neural networks are copied from natural neural systems. Strongly simplified, the present model of neurons consist of two distinct steps in obtaining output from the incoming signals, *i.e.* evaluation and transformation. This is schematically shown in Fig. 1. At the connections between

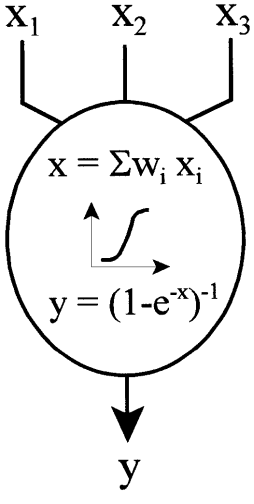


Fig. 1. Scheme of an individual artificial neuron; the input signals x_i were weighted with a factor w_i and then summarized; the output signal y results from a subsequent processing with a transfer function

neurons, the numerical values which represent the input data are multiplied by some 'weights', simulating the different synaptic strength in neural cells. The weight at each of the neurons numerous synapses can have a different value at any given moment. So, the incoming signals can add together to a collective effect, the summarized input signal. Afterwards, the input undergoes an additional nonlinear transformation effected by a so-called 'sigmoidal function' (Fig. 1). However, a single neuron cannot find solutions to complicated applications. Therefore, many neurons must be interconnected as is the case in the brain. The resultant structure is called a neural network (NN). At first the values of the weights in the NN obtained randomly must be corrected by the so-called training of the NN. The most frequently used learning method in the field is the strategy of back-propagation of errors. This is achieved by a defined data set (training data). The main difficulties with performances of NN in solving complex problems are a relevant choice of definition of input and output entities and of the number of hidden layers and of neurons in these hidden layers. General rules do not exist, therefore various possibilities must be tried to find out the best one. Furthermore, the quality of the trained NN must be controlled by another dataset (testing data).

Chemists have made increasing use of these non-linear methods [12]. Some applications dealing with NMR spectroscopy have already been described. So, artificial neural networks have been used for the prediction of ^{13}C NMR chemical shifts of alkanes [13, 14], cycloalkanes [15], trisaccharides [16], monosubstituted benzenes [17], and substituted naphthalenes [18]. Moreover, the cross-peaks in two-dimensional NMR spectra have also been determined with the aim of a NN [19]. Furthermore, the boiling points and inner energies of alkanes have been computed with a NN simulating the relationships between the properties and the ^{13}C NMR spectra of these compounds [20].

In this paper, a neural network for the prediction and the assignment of the ^{13}C NMR chemical shift values of aromatic carbons and its application for the calculation of chemical shifts in polysubstituted benzenes is described.

Results and Discussion

A feedforward multi-layer neural network was used for the present investigation with all neurons in two neighbouring layers connected to each other [21]. In order to take into account the charge distribution over the whole π electronic system, the ^{13}C NMR chemical shifts of all aromatic carbons were calculated by the NN simultaneously. For each of the six aromatic carbons the four substituent induced chemical shift increments I were used for coding the 24 input numbers. Coding of output was also not difficult in our case, because the chemical shift value is a single real number. Consequently, the NN contained 24 input units in the input layer and six output neurons. Only one hidden layer was used. The number of neurons in this layer fixes the number of connections and depends therefore on the problem. An optimized count was found to be 48 in our case. So, the resultant three-layer NN contains about 1500 connections between neurons connected to each other in neighbouring layers. This is shown for some neurons in Fig. 2. Using the simple incremental system, the correlation coefficients between the experimental values of the ^{13}C NMR chemical shifts and the calculated ones were found to be

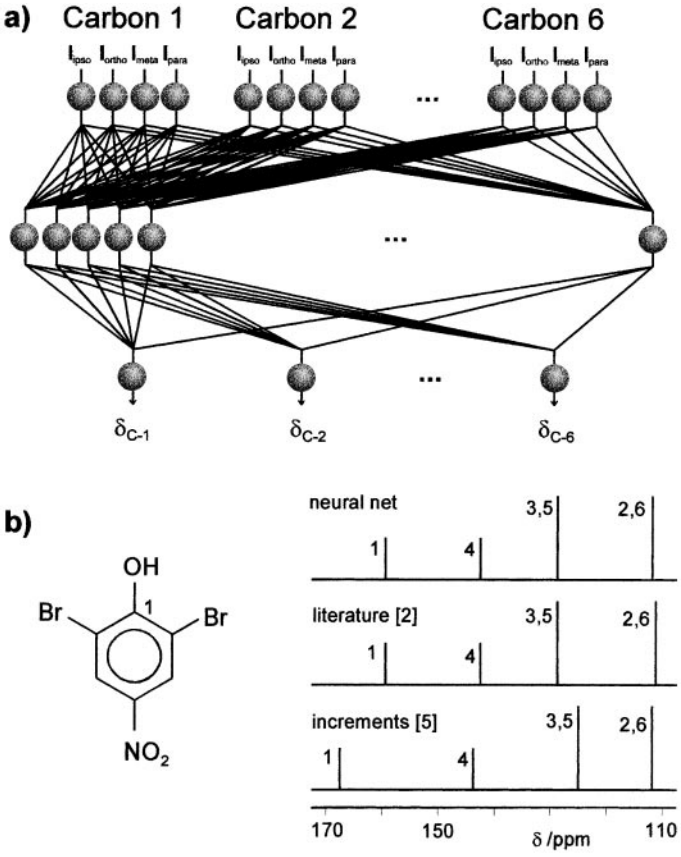


Fig. 2. a) Schematic pattern of the neural network used in this investigation; b) the results computed in this way for 2,6-dibromo-4-nitrophenol compared to the experimental values [2] and to the values achieved with the incremental method

$R_{train} = 0.978$ for the training dataset and $R_{test} = 0.979$ for the test dataset. The comparison of the predicted chemical shift values with the experimental values is shown in Fig. 3a. The mean deviation was found to be 1.5 ppm. These values were improved by the NN to $R_{train} = 0.997$ for the training dataset and $R_{test} = 0.995$ for the test dataset (Fig. 3b). The standard errors decreased to 1.1 ppm for training and test data. The aromatic compounds used here contain one to six substituents. With increasing the number of substituents the chemical shift values predicted by the NN became significant better than the incrementally calculated ^{13}C-NMR chemical shifts. In Fig. 4 the dependence of the average deviation on the substituent number is shown.

Of special interest is the prediction of shift values for compounds where steric effects or hydrogen bonding are effective. Some structures where the consideration of interacting substituents is essential are displayed in Table 1; all calculated and experimentally determined chemical shifts [2] are also given. For 2-ethoxyphenol (**1**), the experimental chemical shift of C-1 is 146.1 ppm. Whereas the NN calculated 145.6 ppm, 141.5 ppm were obtained using increments. In 2,3-dichloroanilin (**2**), the

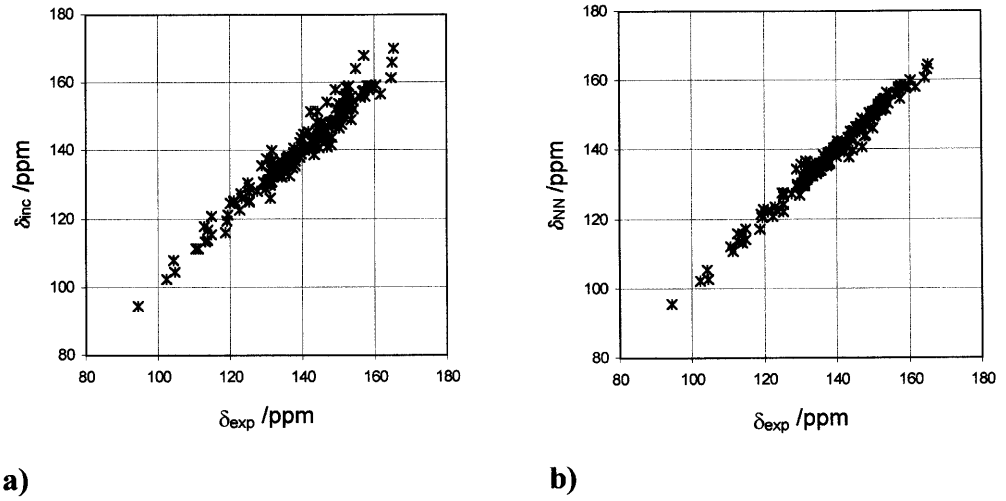


Fig. 3. Correlation of the ^{13}C NMR chemical shift values of the 300 substituted benzenes used in this investigation as a test data set calculated a) by the incremental method ($R = 0.969$) and b) by the neural network ($R = 0.984$) with their experimental values [2]

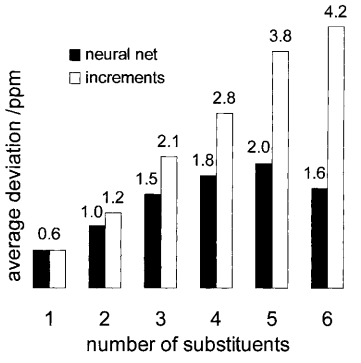
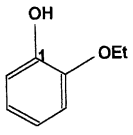


Fig. 4. Comparison of the average deviations which were achieved with the two calculation methods (neural network and incremental system) depending on the number of the substituents; for these calculations, 292 mono-, 891 di-, 401 tri-, 113 tetra- 15 hepta-, and 17 hexasubstituted compounds were used

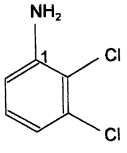
chemical shift of the C-1 carbon (144.7 ppm) was also very well predicted by the NN (144.9 ppm); using increments afforded 148.5 ppm in this case. The chemical shift of C-1 in in 2,4-dimethoxy-5-chloroanilin (**3**) shows a clearly smaller difference by calculation with NN (1 ppm) in contrast to the value calculated using increments (4 ppm). For benzene, which was not a member of the training dataset, the chemical shift value was predicted with high accuracy to be 128.4 ppm. For hexamethylbenzene, the incremental system comes out with an offset of almost 4 ppm, whereas the network predicts the precise value (132.1 ppm).

Table 1. ^{13}C NMR chemical shifts of aromatic carbons of three substituted benzenes calculated by increments and with the neural network; the standard errors (esd) are given in comparison to the experimental values [2]

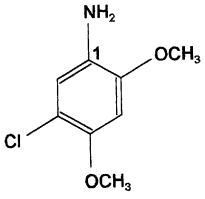
		C-1	C-2	C-3	C-4	C-5	C-6	esd
1	exp.	146.1	146.1	112.2	121.5	120.2	114.9	–
	$calc_{inc}$	141.5	146.3	116.0	112.0	122.3	116.6	5.6
	$calc_{NN}$	145.6	148.0	113.9	120.3	121.2	114.2	1.5
2	exp.	144.7	117.5	133.2	119.5	127.5	113.6	–
	$calc_{inc}$	148.5	121.8	136.0	120.3	128.8	114.6	1.4
	$calc_{NN}$	144.8	120.0	133.7	118.7	127.8	113.5	1.2
3	exp.	131.0	147.7	98.9	146.5	114.0	116.1	–
	$calc_{inc}$	126.0	145.6	101.9	151.3	113.5	117.5	4.0
	$calc_{NN}$	129.8	148.5	102.0	148.7	114.0	116.2	1.7



1



2



3

Conclusions

It is demonstrated that a neural network can learn the association between ^{13}C NMR chemical shifts of aromatic carbons in mono- and polysubstituted benzenes allowing for steric and electronic substituent interactions. Compared to a well-known incremental system, significantly better values were computed with the neural network. Especially for *ortho* and *para* as also for highly substituted benzenes this NN is an effective method for predicting ^{13}C NMR chemical shifts. The average error between the calculated and experimental chemical shift values amounts to 1.1 ppm, *i.e.* it lies within the range of the chemical shift distribution in databases. For this reason, this system is outstandingly suitable for the entry inspection of data material for NMR databases.

Experimental

The NN was trained by a supervised learning method (backpropagation of errors) with about 1000 aromatic structures containing over 200 different substituents [2, 6]. For testing, an independent dataset with about 300 structures was applied. All ^{13}C NMR chemical shifts used here were estimated in $CDCl_3$ or CCl_4 and refer to tetramethylsilane as an internal standard ($\delta_{TMS} = 0.0$ ppm).

References

[1] Gauss J (1995) Ber Bunsengesellsch **99**: 1001
[2] Chemical Concepts, Speclnfo database, STN, Karlsruhe, Germany
[3] NMR Database, Advanced Chemistry Development Inc, Toronto, Ontario, Canada

[4] Robien W, Purtuc V, Schütz V, Felsinger S (1998) CSEARCH. University of Vienna, http://felix.orc.univie.ac.at/wr/csearch_ server_info.html
[5] Pretsch E, Clerc Th, Seibl J, Simon W (1990) Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden. Springer, Berlin Heidelberg
[6] Ewing D (1979) Org Magn Reson **12**: 499
[7] Thomas S, Ströhl D, Kleinpeter E (1994) AROSIM, University of Potsdam, http://www.chem.uni-potsdam.de/arosim/index.html
[8] Gloor A, Cadisch M, Bürgin Schaller R, Farkas M, Kocsis T, Clerc JT, Pretsch E,Aeschimann R, Badertscher M, Brodmeier T, Fürst A, Hediger H-J, Junghans M, Kubinyi H, Munk ME, Schriber H, Wegmann D (1994) SpecTool: A Hypermedia Book for Structure Elucidation of Organic Compounds using Spectroscopic Methods. Chemical Concepts, Weinheim
[9] Hearmon RA, Lin HM, Laverick S, Taylor P (1992) Magn Reson Chem **30**: 240
[10] Fürst A, Pretsch E (1990) Anal Chim Acta **229**: 17
[11] Ströhl D, Thomas S, Kleinpeter E, Radeglia R, Brunn J (1992) Monatsh Chem **123**: 769
[12] Zupan J, Gasteiger J (1993) Neural Networks for Chemists – An Introduction. VCH, Weinheim New York Basel Cambridge Tokyo
[13] Svozil D, Pospichal J, Kvasnicka V (1995) J Chem Inf Comput Sci **35**: 924
[14] Ivanciuc O, Rabine JP, Cabro-Bass D (1998) Comput Chem (Oxford) **21**: 437
[15] El Azzouzi SR, Fan BT, Panaye A, Doucet JP (1997) Organic Reactivity **31**: 3
[16] Clouser DL, Jurs PC (1995) Carbonhydr Res **271**: 65
[17] Slenak S, Kvasnicka V, Pospichal J (1994) Chem Papers – Chemicke Zvesti **48**: 135
[18] Thomas S, Kleinpeter E (1995) J Prakt Chem / Chem Ztg **337**: 504
[19] Corne SA, Fisher J, Johnson AP, Newell W. R (1993) Anal Chim Acta **278**: 149
[20] Meiler J, Meusinger R (1996) In: Gasteiger J (ed) Software-Development in Chemistry, 10. Gesellschaft Deutscher Chemiker, Frankfurt/Main, p 259
[21] WinnNN 0.96, 1995

Received February 8, 1999. Accepted (revised) April 8, 1999