Journal of Magnetic Resonance **157,** 242–252 (2002)
doi:10.1006/jmre.2002.2599

Using Neural Networks for ^{13}C NMR Chemical Shift Prediction–Comparison with Traditional Methods

Jens Meiler,*,[1] Walter Maier,† Martin Will,†,[2] and Reinhard Meusinger‡

*University of Washington, Box 357350, Seattle, Washington 98195-7350; †BASF AG, 67056 Ludwigshafen, Germany; and ‡Universität Mainz, Duesbergweg 10-14, 55099 Mainz, Germany

Received January 29, 2002; revised June 10, 2002

Interpretation of ^{13}C chemical shifts is essential for structure elucidation of organic molecules by NMR. In this article, we present an improved neural network approach and compare its performance to that of commonly used approaches. Specifically, our recently proposed neural network (*J. Chem. Inf. Comput. Sci.* 2000, 40, 1169–1176) is improved by introducing an extended hybrid numerical description of the carbon atom environment, resulting in a standard deviation (std. dev.) of 2.4 ppm for an independent test data set of ~42,500 carbons. Thus, this neural network allows fast and accurate ^{13}C NMR chemical shift prediction without the necessity of access to molecule or fragment databases. For an unbiased test dataset containing 100 organic structures the accuracy of the improved neural network was compared to that of a prediction method based on the HOSE code (*h*ierarchically *o*rdered *s*pherical *d*escription of environment) using SPECINFO. The results show the neural network predictions to be of quality (std. dev. = 2.7 ppm) comparable to that of the HOSE code prediction (std. dev. = 2.6 ppm). Further we compare the neural network predictions to those of a wide variety of other ^{13}C chemical shift prediction tools including incremental methods (CHEMDRAW, SPECTOOL), quantum chemical calculation (GAUSSIAN, COSMOS), and HOSE code fragment-based prediction (SPECINFO, ACD/CNMR, PREDICTIT NMR) for the 47 ^{13}C-NMR shifts of Taxol, a natural product including many structural features of organic substances. The smallest standard deviations were achieved here with the neural network (1.3 ppm) and SPECINFO (1.0 ppm).

Key Words: **neural networks; ^{13}C chemical shift; NMR; method comparison; HOSE code.**

INTRODUCTION

Structure elucidation based on NMR data—manual or automated—can usually be separated into two steps: 1. generation of structural proposals (by hand or with a structure generator), 2. validation of the proposed structures. Based on the ^{13}C NMR spectrum, the latter step requires accurate methods for the prediction of ^{13}C NMR chemical shifts. Considering the huge sets of structures which can be obtained using structure generators (*1–3*), the prediction has to be fast as well, to provide the result in a meaningful time.

The two basic approaches to ^{13}C NMR chemical shift prediction are *ab initio* and empirical calculations. On the one hand, *ab initio* methods can (in principle) calculate the magnetic properties of any molecular structure, such as shielding tensors, shielding anisotropy, and isotropic chemical shifts with respect to an applied magnetic field and the nuclear magnetic moment (e.g., Schindler and Kutzelnigg (*4*), Gauss (*5*) or Cheeseman *et al.* (*6*)). These results can be produced with high accuracy for entire molecular systems starting from an optimized three-dimensional structure of the compound. An important advantage of these methods is that the resulting chemical shift values are not biased by previous experimental results.

However, the necessity to predetermine the correct configuration and conformation in addition to the constitution restricts the applicability of *ab initio* methods. The real three-dimensional structure is often unknown and multiple conformations have to be taken into account for small and flexible molecules, in particular. An extensive optimization of the spatial structure and/or the consideration of multiple conformations make such calculations very time consuming and expensive.

Consequently, the chief advantage of *ab initio* methods is the handling of newly synthesized compounds with exotic structural fragments. Such structures are underrepresented in present databases and their spectrum–structure relationships are therefore often described insufficiently. For example, this was recently shown by Cheeseman and Frisch for cyclopropane, bicyclobutane, [1.1.1]propellane, and oxaspiropentene (*7*).

By contrast, empirical approaches rely on knowledge of chemical shifts from large sets of known molecular structures. The first empirical relationships for ^{13}C chemical shift prediction were given by Grant and Paul (*8*) for paraffins; a few years later, Lindeman and Adams published an extended and convenient correlation chart determined from 59 *n*- and *iso*-alkanes (*9*). Now, the way was open for simple manageable additive methods in chemical shift prediction. The basic idea is the introduction of an individual term (increment) σ_i for each substituent i of a single carbon atom or of a structure fragment (such as, e.g., a benzene ring or a double bond). The values of σ_i were

[1] To whom correspondence should be addressed.
[2] Current address: Aventis SA, 65926 Frankfurt am Main, Germany.

1090-7807/02 $35.00

determined from the observed shift differences between unsubstituted (σ_0) and singly substituted fragments (σ_{i0}):

$$\sigma_i = \sigma_{i0} - \sigma_0. \quad [1]$$

Now, the chemical shift value σ_{calc} is computed by adding all appropriate increments beginning with the value of the nonsubstituted carbon or fragment σ_0,

$$\sigma_{calc} = \sigma_0 + \sum_i \sigma_i + K, \quad [2]$$

where K represents an additional steric correction term.

The advantage of the additivity method is its simplicity, which allows to adopt it easily to other classes of substances. For larger molecules a larger number of increments must be added, so that manual summation becomes tedious. Clerc and Sommerauer presented the first computer program which predicted the ^{13}C NMR shift values for a given chemical structure by means of these additivity rules (*10*). Nowadays computer programs are available to perform this method extremely fast for all ordinary organic substances of relevance. For example, the method of Fürst and Pretsch (*11*) was implemented in the program CS ChemDraw Pro (*12*). Pretsch and co-workers recently presented a new version of their well-known prediction program SpecTool which allows the assignment of both ^{1}H and ^{13}C shift values to a given structure (*13*). However, large deviations between experimental and predicted chemical shift values are often obtained for highly substituted and sterically hindered structures. A limitation of this simple approach is caused by the disregard of the intramolecular interactions between individual substituents. Therefore so-called cross terms σ_{ij} were introduced into Eq. [2] to compensate for these imperfections. However, to consider all possible interactions between three and more substituents, the introduction of individual cross terms ($\sigma_{ijk}, \sigma_{ijkl}, \ldots$) for every single case would be needed, which quickly becomes difficult to parameterize reliably.

Since the beginning of NMR spectroscopy, experimental data have been collected in special libraries, together with the corresponding structures. Simultaneously with the enhancement of computer technology these data were stored in spectra databases (*14*). In addition to the relatively simple storage of ^{13}C NMR chemical shift values, the one-dimensional coding of the chemical environment of each carbon atom also became necessary for efficient electronic data processing. The *h*ierarchically *o*rdered *s*pherical description of *e*nvironment (HOSE) code, developed by Bremser (*15*), is suitable for this problem. Starting from the atom of interest, all atoms bonded directly to this atom (1st sphere), over two bonds (2nd sphere), and so on are coded using characters which define atom types, bond types, ring closures, and spheres. The number of described spheres depends on the length of the code.

The prediction of ^{13}C NMR chemical shifts was possible for any molecular structure by computing the HOSE code for each carbon and subsequent similarity search in the database. Provided that the database contains similar HOSE codes, this method allows an exact prediction of carbon chemical shifts. However, the accuracy of the prediction strongly depends on the similarity between the new and known HOSE codes and on the quality of the database (e.g., number of wrongly assigned spectra or systematic errors from the experiment). Here a further advantage of databases becomes obvious—the reference to all original data. This enables a supplementary cross-check of the assignment. Disadvantages of this method are the relatively long CPU time (~1–60 s depending on size and accessibility of the database) compared with additivity methods (few ms), the enhanced uncertainty for structures outside the covered structural space in comparison with quantum chemical approaches, and the necessity of access to large databases. Existing implementations of such databases for ^{13}C NMR chemical shift prediction are SpecInfo (*14*), ACD/CNMR (*16*), and PredictIt NMR (*12*).

With the introduction of artificial neural networks into chemistry in recent years (for an introduction to the field read, e.g., (*17*)), their potential for ^{13}C NMR chemical shift prediction was evaluated. Similarly to the historical development of incremental methods, they were applied first to several classes of substances (*18–24*). Applications covering the overall space of described organic molecular structures were published in last years by Robien and LeBret (*25, 26*). Recently, we have shown the introduction of the numerical HOSE code description into a neural network for the same purpose (*27*).

In this paper we describe an essential enhancement of this neural network approach and compare the model with a variety of traditional methods such as additivity rules, fragment-based database searches, and quantum chemical calculations to review the existing implementations as well as their advantages and disadvantages.

RESULTS AND DISCUSSION

Computation of ^{13}C NMR chemical shifts with an artificial neural network requires the numerical description of the chemical environment of the carbon atom of interest. This description must meet certain conditions to be suitable as input for a neural network. Thus, the input vector should be constant in length and independent of the number of descriptors necessary for coding the structural environments. Further, a particular digit of the vector should always describe the same structural feature, independent of the actually coded carbon atom. Only the chosen value applied on an input is influenced by the specific description of the structural environment of the actual carbon atom. These conditions meet the requirements for reasonable training of the neural network, since the influence of a respective structural feature can be understood by a certain input unit.

In consequence, every constitutionally different atom position around a carbon needs individual input units. The maximum number of potential atom positions of one substituent depends on the number of considered spheres (compare Fig. 1). Only one position must be considered in the first sphere, up to three

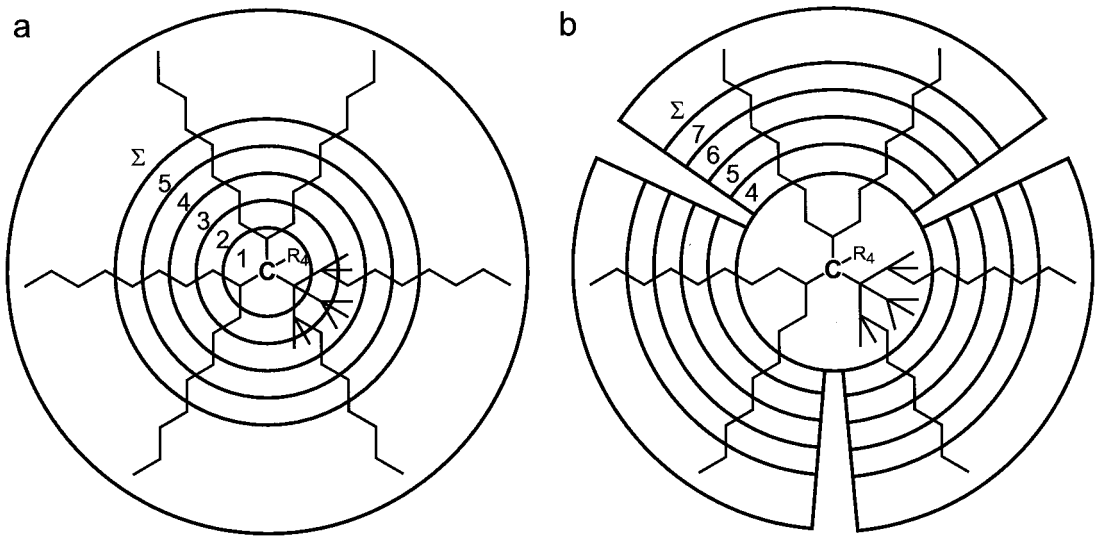

FIG. 1. Comparison of the description of atom environment used in the first version of C_SHIFT (a) with the newly introduced description (b). The number of spheres considered was increased from 5+ sum sphere to 7+ sum sphere. The individual substituents of a carbon atom form individual parts of the description in the new version. All atoms in the first three spheres around an observed carbon atom are described by eight atom parameters.

possible positions in the second sphere, up to nine in the third, and up to 2187 in the eighth sphere. Thus, to code the environment of a quaternary carbon atom over eight spheres,

$$4 \times (1 + 3 + 9 + 27 + \cdots + 2187) = 13,120 \qquad [3]$$

descriptors would be necessary. This number is unacceptably large for the input layer of a neural network, given the resulting extremely large number of weights and the unreasonably long training time. This large number of input neurons may be reduced using so-called sum parameters that describe not only a single atom, but the sum obtained for a group of atoms, e.g., all atoms in a specific sphere. Especially, the large number of possible atom positions in the outer spheres, which are obviously rarely sampled, is drastically reduced by this procedure. All atoms that occur in organic substances (and are therefore frequently sampled in ^{13}C NMR databases) are subdivided into different atom classes by analyzing their atomic number, hybridization state, and number of bound hydrogen atoms by a straightforward approach (compare, e.g., Table 1 or (*27*)). The environment of a carbon atom can now be assessed by evaluating the number of occurrences of each of these atom classes in every sphere. Therefore, the number of necessary input neurons is reduced to the number of atom classes multiplied by the number of outer spheres. Considering, e.g., 30 atom classes and six spheres without the distinction of different substituents, only $6 \times 30 = 180$ would be necessary.

Such an approach was introduced (*27*) by defining 28 atom classes and two additional sum parameters counting the number of hydrogen atoms and the number of ring closures in each sphere. These 30 numbers were collected for the first five spheres plus one additional sum sphere that contained all atoms of the sixth and higher spheres, yielding 180 input parameters. To

TABLE 1
Introduced Classes of Atoms and the Number of Atoms in Each Class in the Used Set of Data

ID	Atom class	Frequency of atoms in this class
1	⟩C⟨	54 074
2	⟩CH−	134 203
3	−CH_2−	280 331
4	−CH_3	202 109
5	=C⟨	158 815
6	=CH−/=CH_2	77 808
7	≡C−/≡CH/=C=	10 119
8	⟩⟩C− (aromatic)	195 095
9	⟩⟩CH (aromatic)	302 212
10	⟩N−	32 238
11	−NH−	28 411
12	−NH_2	7 917
13	=N−/=NH	26 842
14	≡N	5 905
15	−NO_2	6 140
16	⟩⟩N (aromatic)	12 966
17	=N≡	423
18	−O−	100 359
19	−OH	59 877
20	=O	112 320
21	⟩P−/−PH−/−PH_2	631
22	⟩PO−	2 220
23	−S−/−SH	11 358
24	=S	3 264
25	⟩S=O	740
26	⟩SO_2	5 307
27	−F	15 862
28	−Cl	24 454
29	−Br	6 109
30	−I	1 132
31	Sum of all hydrogen atoms bound in this sphere	
32	Sum of all ring closures in this sphere	

consider the additional influence of a conjugated π-electron system, a second vector of 180 numbers was determined, in which only atoms were counted that belong to the same π-electron system as the carbon atom of interest. Thus, the three-layer neural network contained 360 input units. For each of the nine individual carbon atom classes used here, an individual neural network was trained. The average deviation for predicting the chemical shift values of an independent validation set of about 1000 molecules with more than 15,000 carbons was 1.8 ppm (*27*). This result comes close to the deviations obtained by the more time-expensive HOSE code predictions. However, larger deviations were obtained for conjugated systems, tertiary and quaternary carbon atoms, small ring systems, and double-bonded carbon atoms. This weakness of the neural network compared to databases was discovered by the chemical shift prediction of marine natural compounds which were computed previously by the structure generator CoCon (*28*). To address this shortcoming the description of the atom environment was modified as shown in Fig. 1.

The number of spheres considered in the code was increased from 5 plus the sum-sphere to 7 plus the sum-sphere. This modification ensured that the influence of a conjugated π-electron system on the chemical shift was felt even more. To improve the prediction of quaternary and tertiary carbon shifts the individual substituents of a carbon atom were handled separately; an input vector is computed and input neurons are provided for each substituent of the carbon. Thus, possible interactions between these substituents are described more precisely than in the original code. Consequently, the numerical description of a quaternary carbon atom is four times longer than a methyl group description. Individual neural networks are trained for primary, secondary, tertiary, and quaternary carbon atoms with well-defined but different sizes of the input layer. A further modification is suggested by the problems that occur in predicting the chemical shift for small ring systems, conjugated π-electron systems, and highly substituted carbon atoms. The important interactions and influences of the inner sphere atoms are not properly picked up by the sum parameters. Since the number of necessary descriptors for an individual coding of atoms stays rather small for inner spheres, these atoms were no longer described by sum parameters but rather coded individually by introducing eight parameters: number of valence electrons, period, electronegativity, van der Waals radius, hybridization, bond type to previous atom, number of bonded hydrogen atoms, and ring closures. Further, the number of atom classes was enlarged to 30 to include azides −N═N≡N and sulfoxides ⟩S═O (Table 1, ID 17 and ID 25). Consequently, the resulting code of a carbon substituent has been lengthened to 13×8(inner spheres) $+ 64 \times 5$(outer spheres) $= 424$ descriptors. This is schematically shown in Figs. 1 and 2.

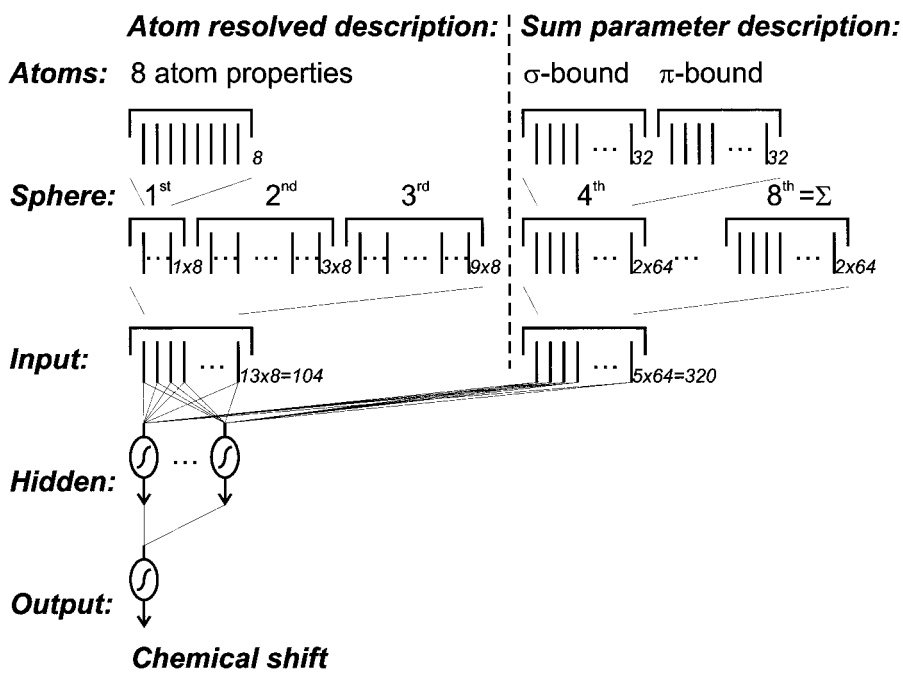

FIG. 2. The description of the atom environment is shown by the example of one single substituent on an "atom" level and on a "sphere" level. Further, the architecture of the neural networks is given. An "atom resolved description" is used in the first three spheres, which consists of 8 parameters per atom. For each of the next four spheres and the additional sum-sphere a "sum parameter description" is applied. The occurrences of the 32 defined atom types are counted two times, the first time considering all atoms, the second time considering only atoms that belong to the same π-electron system as the atom of interest. The atoms in the first three spheres (up to 13) are treated by eight input neurons each, which leads to 104 input units. The 64 input values necessary to code one sphere lead to 320 parameters for spheres 4, 5, 6, 7, and the sum-sphere. This input vector is applied for *every* substituent of a carbon atom, which leads to a varying number of input units (Table 2). The data are processed by a hidden layer and one output neuron computes the chemical shift value. The number of hidden neurons is 8 for allyl carbon atoms (class 7) and 32 in all other cases.

TABLE 2
Results for Training and Testing the Artificial Neural Networks

			Training and monitoring data			Test data		
ID	Atom type	Net structure inp-hid-out[a]	Count	std. dev. [ppm]	aver. dev. [ppm]	Count	std. dev. [ppm]	aver. dev. [ppm]
1	⟩C⟨	1 696-32-1	52 379	1.9	1.2	1 695	2.9	1.7
2	⟩CH−	1 272-32-1	130 289	2.6	1.8	3 914	3.1	2.1
3	$-CH_2-$	848-32-1	272 189	2.1	1.4	8 142	2.4	1.5
4	$-CH_3$	424-32-1	195 842	1.9	1.1	6 267	1.9	1.2
5	=C⟨	1 272-32-1	153 952	2.7	1.8	4 863	3.4	2.2
6	$=CH-/=CH_2$	848-32-1	75 415	2.3	1.6	2 393	3.1	2.1
7	≡C−/≡CH/=C=	848-08-1	9 837	2.2	1.4	282	3.0	1.9
8	))C− (aromatic)	1 272-32-1	189 197	2.2	1.5	5 898	2.5	1.7
9	))CH (aromatic)	848-32-1	293 169	1.7	1.1	9 043	1.7	1.1
Weighted average			1 372 269	2.1	1.4	42 497	2.4	1.6

[a] Structure of the used three-layer neural network : number of input units – number of hidden neurons – number of output nuerons.

Overall, a set of 1.3 million descriptions was constructed out of the SpecInfo database (*14*). To consider the carbon atom classes listed in Table 2, nine different neural networks were trained using the program Smart (*29*). For this, the available data material was subdivided into three groups. The predominant majority (∼95% of all molecules) was employed for the training of the networks whereas the remaining molecules served as a monitoring set (∼2% of all molecules) and as an independent test set (∼3% of all molecules). The monitoring set provides the precise moment of interrupting the training process to prevent so-called overtraining. The test set serves as an independent validation of the final neural networks. All results are summarized in Table 2. In Fig. 3 the quality of the prediction is illustrated by two correlation diagrams for the training and the test set of molecules, respectively. The trained neural networks are implemented in a new version of the C_Shift software (*27, 29*). Only a slightly better average deviation (aver. dev.) of 1.6 ppm was obtained in comparison to the former neural network (aver. dev. = 1.8 ppm) (*27*). Nevertheless, tremendous improvements are obtained for some of the discussed structural fragments, as explained in the following paragraphs.

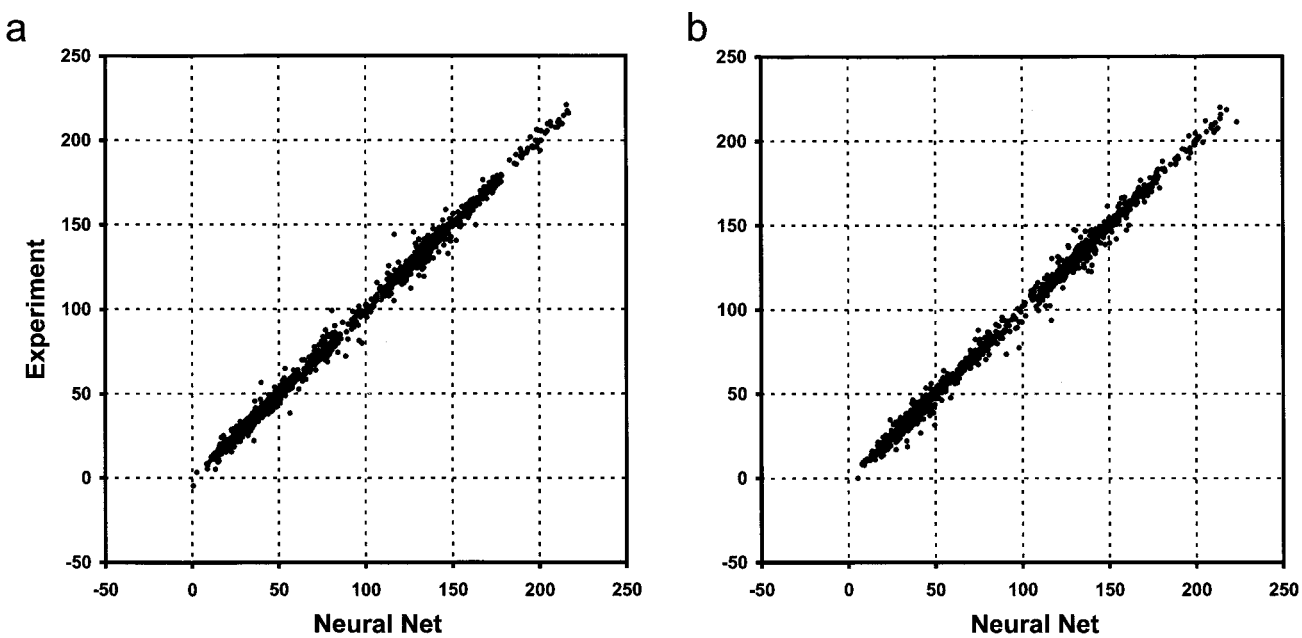

FIG. 3. Correlation diagrams illustrating the results obtained by the neural network chemical shift prediction for the training set of data (a) as well as for the test set of data (b). In both diagrams ∼2000 randomly selected carbon atoms are visualized.

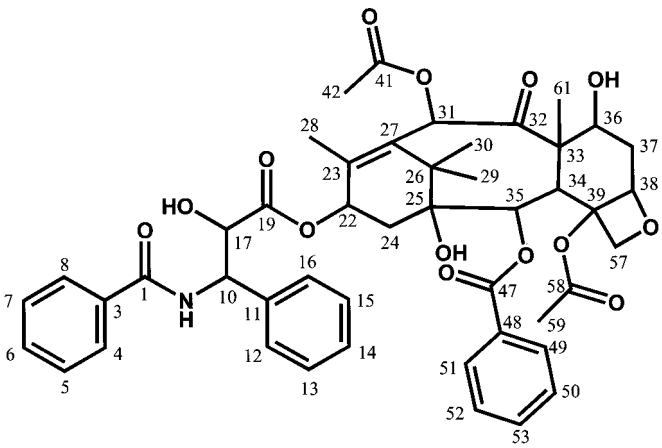

FIG. 4. Constitution of the anti-tumor agent Taxol used to compare the neural network prediction of chemical shifts with other methods. The unusual order of numbering the atoms was introduced by the SpecInfo database. Stereochemistry is only considered by the Gaussian and Cosmos programs. All other prediction methods rely on this constitution formula only.

A dramatic decrease of the already small average deviation was not expected based on these modifications. A decrease of the maximum deviations for carbon atoms involved in the mentioned structural features is useful, especially for the prediction of structures outside the structural space covered by databases.

A detailed comparison of different chemical shift determination methods is performed by analyzing experimental and predicted chemical shift values of the antitumor agent Taxol (*30*) (Fig. 4). Multifaceted structural features common in organic chemistry are to be found in this natural compound. The chemical shifts of the 47 carbons were available both as experimental data (*31*) and as results of *ab initio* calculations (*6, 7*). The experimentally determined ^{13}C NMR shift values were compared with the calculated ones from both versions of the C_Shift program, with the HOSE code prediction by the SpecInfo database (*14*), with the Bio-Rad PredictIt NMR tool (*32*), and with the ACD software package CNMR (*16*). A comparison was also made with results from the incremental methods CS ChemDraw Pro (*12*), and SpecTool (*13*), the *ab initio* calculations from Gaussian 98 (*33*) (the chemical shifts were computed from NMR calculations performed at the HF/6-31G(d) level using STO-3G optimized geometries), and a quantum chemical calculation with the Cosmos 4.5 program (*34*). (In this case, three C-atoms were excluded since no suitable groups were available for the parameterization. Chemical shifts are predicted on the basis of their proportional behavior with respects to bond polarization energy utilizing bond polarization theory (BPT) (*35–37*)).

Table 3 presents the results for all these different methods for the 47 individual carbon atoms. The deviations of the experimental chemical shift values are represented graphically by the area of the circle next to the numerical value. This allows fast comparison of the different procedures (columns) and of the individual structural features (rows). In addition, the standard deviation (std. dev.),

$$\text{std.dev.} = \sqrt{\frac{\sum_{i=1}^{n} \left(\sigma_i^{\text{estim}} - \sigma_i^{\text{exp}}\right)^2}{n-1}}, \qquad [4]$$

is given for the different methods to estimate the error for an unknown carbon.

On closer examination of values in Table 3, it is striking that the accuracy of the prediction varies between different methods as well as between different atoms. A continuous enlarged inaccuracy is shown for the aromatic carbon atoms C-3, C-6, C-11, and C-14 with all methods. This suggests that the experimental values themselves have been wrongly assigned for the two *ipso*-carbons C-3 and C-11 and the two *para*-carbons C-6 and C-14, respectively. After the experimental values are exchanged, the average deviations of the computed values from these four atoms decrease from 5.2, 3.1, 4.5, and 4.3 to 0.8, 0.5, 0.1, and 0.7 via all nine prediction methods. In what follows, we assume that these experimental values should be exchanged. In Table 3 the updated overall std. dev., aver. dev. and maximal deviation values (if necessary) are given in brackets.

The HOSE code prediction method of the SpecInfo database achieves a std. dev. of 1.0 ppm using five spheres for the description of the carbon atom environment. To allow an unbiased comparison Taxol was removed from the database prior to predicting the chemical shift values. The neural network that uses the newly introduced description method comes very close to these excellent results by achieving a std. dev. of 1.3 ppm. Also in this case Taxol was not part of the dataset that was used for training the network connections. The maximal obtained deviation for a single carbon atom is very small with 3.4 ppm for the neural network and 3.0 ppm for the HOSE code prediction. The improvement of the new neural network becomes obvious by comparing it with the former approach; the standard deviation is less than half of that with the previous neural network, and the maximal deviation also drops by more than 50%. This result is due to the fact that some large deviations for single carbon atoms (e.g., C-10, C-23, C-36, C-38, C-57) are dramatically reduced. These carbon atoms generally belong to one of the earlier mentioned structural features that are insufficiently described in the old method: C-38 and C-57 belong to a four-membered ring, C-23 participates in a double bond, and C-36 and C-10 are tertiary carbon atoms. PredictIt NMR achieves a std. dev. of 3.7 ppm, again after removing Taxol itself and similar structures from the fragment database. The largest deviations are obtained for the sterically hindered ring system C-32, C-33, C-34, C-35, C-36, C-37, C-38, C-39, and C-57. The second problematic region is also situated in the ring system at C-22, C-23, C-27, and C-31. Here the lack of fragments that describe such structural features becomes obtainable. The remaining chemical shifts are predicted remarkably well. The CNMR package achieves a std. dev. of 2.9 ppm. Even larger deviations are observed here for the fragments C-22,

TABLE 3
Comparison of Experimental and Predicted ^{13}C Chemical Shifts for Taxol

		Neural networks		Increments		HOSE code predictions			Quantum chemical	
ID	Exp.	*C_Shift (new)*	*C_Shift (old)*	*CS Chem Draw Pro*	*SpecTool*	*SpecInfo*	*Predict It NMR 1.3*	*CNMR 6.0*	*Cosmos 4.5*	*Gaussian 98*
1	167.0	166.1	167.2	167.9	167.6	166.8	167.6	166.6	157.7	190.6
3	138.0	*136.0*	*133.9*	*133.5*	*133.5*	*133.2*	*134.2*	*136.3*	*121.7*	*135.7*
4	127.0	127.8	127.4	127.3	127.3	127.6	127.7	127.2	126.8	133.6
5	129.0	128.6	129.0	128.6	128.6	128.1	128.4	128.1	130.8	124.5
6	128.3	*131.9*	*132.3*	*131.9*	*131.9*	*131.6*	*132.0*	*131.7*	*129.6*	*132.9*
7	129.0	128.6	129.0	128.6	128.6	128.1	128.4	128.1	131.3	125.6
8	127.0	127.8	127.4	127.3	127.3	127.6	127.7	127.2	128.1	133.6
10	55.0	55.0	59.9	52.4	51.6	54.6	54.9	55.8	54.6	59.1
11	133.6	*136.7*	*138.1*	*142.4*	*138.8*	*136.7*	*140.7*	*136.1*	*135.8*	*138.7*
12	127.0	126.8	128.0	127.1	128.3	127.2	126.9	128.3	128.0	130.1
13	128.7	128.3	130.0	128.3	128.6	128.6	128.4	129.2	130.1	127.8
14	131.9	*127.9*	*127.7*	*126.5*	*125.8*	*127.7*	*127.4*	*126.2*	*130.6*	*128.2*
15	128.7	128.3	130.0	128.3	128.6	128.6	128.4	129.2	129.4	127.6
16	127.0	126.8	128.0	127.1	128.3	127.2	126.9	128.3	129.9	130.7
17	73.2	73.3	70.8	85.2	85.2	72.9	74.0	74.0	75.4	76.0
19	172.7	171.4	171.7	172.0	172.0	172.8	172.1	172.3	163.8	178.6
22	72.3	71.5	70.8	71.1	73.1	72.0	75.8	64.3	70.5	69.2
23	142.0	140.6	130.4	132.9	132.2	139.4	134.0	132.8	142.3	137.0
24	35.7	33.9	33.2	34.9	34.9	35.8	41.1	36.3	35.8	38.6
25	79.0	79.3	82.6	81.3	81.3	78.7	79.7	77.4	77.4	73.1
26	43.2	43.7	44.3	33.4	35.4	42.8	41.5	44.8	47.2	41.0
27	133.2	133.2	131.0	138.5	141.6	134.6	137.1	142.7	118.7	133.8
28	14.8	16.9	18.4	10.9	12.9	16.3	16.6	19.8	-	18.5
29	26.9	24.1	23.1	16.3	16.3	24.0	21.4	29.3	22.4	26.7
30	21.8	24.1	23.1	16.3	16.3	24.0	21.4	26.8	21.9	24.5
31	75.5	77.9	73.2	79.6	81.6	76.0	74.0	66.9	76.9	80.0
32	203.6	204.7	207.3	212.8	213.9	204.4	210.2	205.4	-	206.4
33	58.6	59.6	56.9	39.6	41.6	57.9	55.3	58.7	54.5	54.3
34	45.6	44.3	43.4	27.7	29.7	44.3	52.4	45.6	48.4	44.7
35	74.9	74.1	72.7	70.0	70.0	74.2	80.7	75.6	73.4	71.9
36	72.2	74.9	66.8	64.9	66.6	73.3	75.4	73.2	69.8	65.5
37	35.6	33.1	31.9	28.9	30.9	36.6	36.5	36.0	37.4	30.6
38	84.4	82.0	74.9	86.7	86.7	81.9	74.0	84.2	75.4	82.9
39	81.1	82.3	80.5	79.2	79.6	81.1	77.0	81.2	76.5	79.8
41	171.2	170.3	169.6	171.0	171.0	170.6	170.1	168.9	174.9	176.7
42	20.8	21.5	20.3	27.4	17.4	21.7	21.0	21.0	22.4	26.2
47	167.0	166.5	167.7	167.0	167.0	166.5	165.8	166.8	159.4	173.0
48	129.1	129.2	128.4	130.5	130.5	129.4	130.7	129.3	128.7	128.1
49	130.2	129.8	129.0	129.7	129.7	129.5	129.7	130.3	124.9	130.4
50	128.7	128.3	129.1	128.4	128.4	128.9	129.1	128.7	130.2	126.1
51	130.2	129.8	129.0	129.7	129.7	129.5	129.7	130.3	124.5	135.3
52	128.7	128.3	129.1	128.4	128.4	128.9	129.1	128.7	130.2	127.8
53	133.7	132.8	133.4	132.8	132.8	132.8	132.8	133.9	130.4	136.9
57	76.5	75.7	72.6	79.6	79.2	76.0	65.0	76.7	69.2	73.1
58	170.4	171.1	170.0	171.0	171.0	170.5	170.2	170.5	171.7	175.8
59	22.6	21.7	21.2	17.9	17.9	22.0	21.6	22.6	21.4	26.9
61	9.5	12.9	16.6	10.3	10.3	11.6	17.0	10.7	-	21.2
std. dev.:		1.6(1.3)	3.3(3.1)	5.8(5.6)	5.4(5.3)	1.5(1.0)	3.9(3.7)	3.1(2.9)	4.6(4.3)	5.2(5.1)
aver. dev.:		1.2(1.0)	2.4(2.1)	3.8(3.5)	3.7(3.3)	1.0(0.8)	2.7(2.3)	1.7(1.6)	3.4(3.3)	4.0(3.9)
max. dev.:		4.1(3.4)	11.6	19.0	17.0	4.8(3.0)	11.5	9.5	16.3(14.5)	23.6

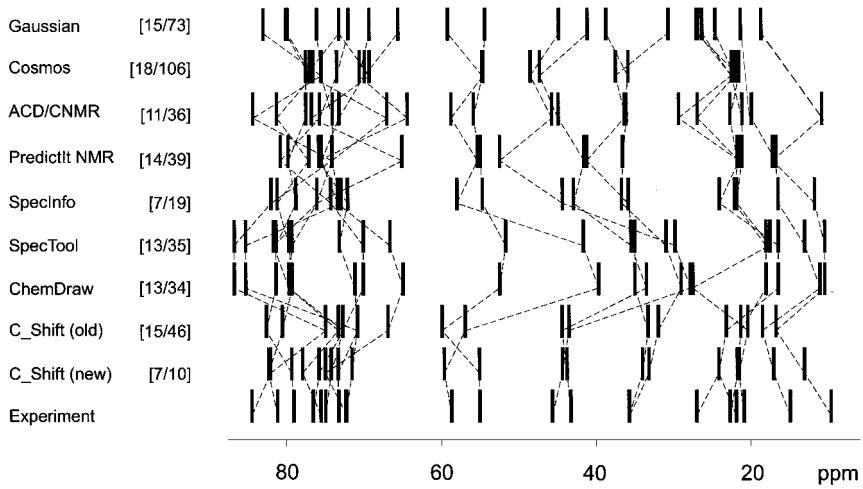

FIG. 5. Comparison of the experimentally obtained and several predicted ^{13}C NMR spectra. The number of incorrect assignments that would be obtained by ordering the carbon atoms just by the corresponding predicted chemical shift value is given by the first integer in the brackets. The second integer holds the total number of inversions—all cases in which the relative order of two predicted chemical shifts is the opposite of that observed experimentally.

C-23, C-27, C-31. However, in this case Taxol itself and also 14 similar structures are part of the ACD database and therefore also part of the HOSE code library (which of course makes it easy to predict chemical shifts very exactly). The large deviations obtained for the mentioned region are probably caused by the variance of the 15 structures in the substituent attached to C-31 and its enormous influence on the attached double bond.

Incremental methods (std. dev. 5.6 ppm and 5.3 ppm), as well as quantum chemical calculations (std. dev. 4.3 ppm and 5.1 ppm), gave worse results. Incremental methods fail for highly substituted carbon atoms with interacting substituents such as C-33 and C-34. While the quantum chemical calculation supplied good results here, unexpectedly large deviations appear in the case of C-1 and C-61. Further deviations were found for a multiplicity of carbons independent of their hybridization state and degree of substitution, e.g., C-4, C-8, C-19, C-23, C-36, C-42, C-47, and C-58 in the case of GAUSSIAN and C-3, C-19, C-27, C-38, C-47, C-49, C-51, and C-57 for COSMOS. This uncertainty is often in contradiction to the weaknesses of the other methods. In such a way, the quantum chemistry calculations supplement database methods.

The GAUSSIAN and COSMOS programs start from a three-dimensional structural model. All other implementations use exclusively the constitution of the molecule for the chemical shift prediction and are unable to distinguish between the diastereotopic atoms C-29 and C-30. Thus, the prediction of ^{13}C chemical shift cannot be perfect for every atom. (ACD reports two different values for the methyl groups only because the experimental spectrum is in the database.) Neglecting the three-dimensional structure can lead to single deviations of up to 30 ppm in some rare cases. However, usually the deviation that is introduced by this simplification lies well below 5 ppm. The error resulting from experimental uncertainty is suggested to lie between 0.5 and 1.0 ppm, comparing ^{13}C NMR spectra obtained in different experiments but using the same solvent and working at the same temperature.

Figure 5 provides a visual comparison of the obtained ^{13}C NMR spectra. Since often not the absolute deviation in predicting ^{13}C NMR chemical shifts is important, but the possible incorrect assignments caused by these deviations, we count the number of incorrect assignments if the carbon atoms are ranked by their predicted chemical shift values (Fig. 5). This number is 7 for the SPECINFO, and the new implementation of the neural networks is in the range of 11 to 18 for the CNMR, PredictIt NMR, the old implementation of the neural network, both incremental methods, and both quantum chemical methods. The total number of inversions (inversion means that the relative ranks of two predicted shifts is the opposite of that observed experimentally) is a more sensitive measure of the probability of incorrect assignments, since every single inversion would be an incorrect assignment in a molecule with just these two carbon atoms. The maximal number of possible inversions is therefore $\frac{1}{2} \times 46 \times 47 = 1081$ for Taxol. The value is especially sensitive for large deviations in some of the predicted shifts since they lead not only to one but several inversions.

For this benchmark, the new implementation of the neural networks is better than SPECINFO, having only 10 inversions (0.9%),

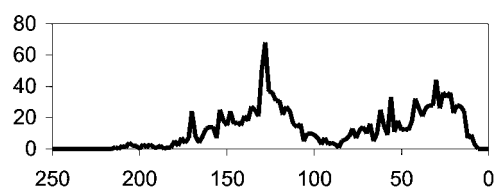

FIG. 6. Distribution of the 1547 ^{13}C NMR chemical shifts in the 100 oragnic molecules used to compare the advanced neural network implementation with the former implementation and with SPECINFO HOSE code prediction.

TABLE 4
Results of Chemical Shift Prediction of 1547 Carbons in 100 Newly Formed Structures, Which Were Not Included in the Database

		C_Shift (new)			C_Shift (old)			SpecInfo		
ID[a]	Count	std. dev.	aver. dev.	max. dev.	std. dev.	aver. dev.	max. dev.	std. dev.	aver. dev.	max. dev.
1	68	3.8	2.5	16.0	3.7	2.7	16.8	1.9	0.9	10.7
2	149	3.1	2.3	11.8	4.4	3.4	13.2	2.9	1.7	12.2
3	288	2.4	1.8	8.9	2.9	2.1	10.8	2.5	1.4	15.4
4	242	2.0	1.3	7.6	2.3	1.5	10.0	1.5	0.9	7.3
5	190	3.3	2.3	12.3	5.3	3.6	22.8	2.7	1.5	21.9
6	118	2.9	2.2	10.1	5.6	4.1	25.0	2.3	1.4	15.7
7	12	3.2	2.5	7.2	4.7	4.0	7.8	3.6	2.7	7.3
8	233	2.8	2.1	10.7	3.9	2.8	15.6	3.6	2.1	18.2
9	247	1.7	1.2	5.3	2.0	1.4	8.1	2.2	1.2	11.8
All:	1 547	2.7	1.9	16.0	3.7	2.5	25.0	2.6	1.4	21.9

[a] Nine different carbon atom classes (see Table 2).

compared with 19 inversions (1.7%) for SpecInfo. This value is significantly higher for ChemDraw, SpecTool, PredictIt NMR, and CNMR with 34, 35, 39, or 36 (~3%) inversions, respectively, as well as for the first implementation of the neural networks (46 inversions, 4.2%). Gaussian and Cosmos achieve the highest values with 73 (6.8%) and 106 (9.8%) because of the large systematic errors for some of the predicted chemical shift values.

The accuracy of the two quantum mechanical methods is worst for our example molecule and lies below expectations. For small molecules where a precise geometry is available such methods are known to predict the ^{13}C chemical shift values with deviations as low as 1–2 ppm (*6*). The significantly higher deviations obtained here are mainly caused by the inaccuracy of the three-dimensional model (which is hard to compute precisely for such "large" molecules) rather than by the accuracy of the chemical shift prediction itself.

A statistically more valid test of the method is performed using a small database of 100 molecules that are not part of the SpecInfo database and therefore also not used at all for training the neural network. These substances reflect a broad range of organic chemistry, as shown by plotting the distribution of the chemical shift values in Fig. 6 (the distribution is very similar to the distribution of all chemical shifts in the SpecInfo database). The chemical shift values of ~1500 carbon atoms are computed by C_Shift and the SpecInfo database HOSE code prediction only. The deviations to the experimental values are compared for all nine introduced carbon atom classes in Table 4. The improvement for classes 2, 5, 6, 7, and 8 by switching to the newly introduced atom environment description is clearly to be seen. The method approaches the result yielded using a HOSE code prediction with the complete SpecInfo database in the background, although slight differences have to be mentioned: The networks are slightly better for the aromatic carbon atom classes 8 and 9 since conjugated systems are certainly better described in the introduced description compared to the HOSE code prediction. The neural networks produce slightly worse results for the classes 1, 5, and 6. Obviously, the large number of input neurons necessary for coding the substituents in comparison to the relatively low number of the corresponding carbon atoms in the training set of data lead to a slightly worse model in these cases. This is suggested by the large differences between the deviations of the training and the test data sets for these classes (cf. Table 2). However, the std. dev. of the test data set could not be decreased by reducing the number of neurons in the hidden layer.

Overall the artificial neural networks approach the accuracy of HOSE code prediction methods in quality. However, keeping in mind that the trained neural networks are database-independent and about 1000 times faster than the HOSE code predictions, the method is well suited for routine work on a PC and for the fast screening of large numbers of molecules.

CONCLUSION

In this paper a new numerical description for the constitutional environment of a carbon atom is introduced as input for artificial neural networks to predict the ^{13}C chemical shift. Artificial neural networks are capable of combining the twin advantages of increments and HOSE code chemical shift prediction: rapidity and accuracy. After expensive training to be carried out only once, they run a prediction about 1000 times faster and independent of direct access to a database. Since interactions between substituents are considered in the training data, they are more precise than incremental methods. They are also expected to be more accurate in comparison to the HOSE code prediction and incremental methods in estimating chemical shifts of newly synthesized molecules which are badly represented in databases.

Based on nine different neural networks, the C_Shift program is able to predict the ^{13}C NMR chemical shift of all organic compounds that contain exclusively H, C, N, O, P, S, and the halogens. The method achieves a standard deviation (std. dev.) of 2.4 ppm and 2.7 ppm for two sets of independent test molecules,

respectively. These results are comparable with the results obtained using a five-sphere HOSE code prediction relying on the SpecInfo database (std. dev. = 2.6 ppm).

By the example of the antitumor agent Taxol as an organic compound including many typical structural features, the capability of the neural networks is shown in comparison to a broad variety of prediction methods. The best results were obtained using SpecInfo HOSE code prediction and the here introduced artificial neural networks (std. dev. = 1.3 ppm). The HOSE code predictions of PredictIt NMR and ACD CNMR achieve std. dev. of 3.7 ppm and 2.9 ppm, respectively. However, the availability of Taxol fragments in the ACD database biases this result. Incremental methods as ChemDraw and SpecTool supply worse results for this complicate structure (std. dev. 5.6 ppm and 5.3 ppm). The much more extensive quantum chemistry routines of Cosmos and Gaussian lead also to rather large deviations (std. dev. 4.6 and 5.1 ppm), mainly caused by incorrect representation of three-dimensional structure and dynamics rather then by the prediction method itself.

Methods that rely on fragment databases (HOSE codes) are fully capable of incorporating new structures fast into their prediction by adding the fragments or HOSE to their database. Also neural networks could be retrained dynamically to be adjusted if new substances of even novel substance classes are found. However, *ab initio* methods do ideally not need any adjustment at all in predicting new classes of substances.

EXPERIMENTAL

All neural networks were created, trained, and analyzed using the software Smart (*29*). The three-layer networks were trained using back propagation of errors (e.g., (*17*) or (*38*)). The learning rate η was decreased from 0.01 to 0.0001 during the training procedure and the momentum α was set to 0.5. The training took between ~300 and ~7000 iterations, depending on the size of the network and the training set of data. The optimization was performed on a O2 workstation equipped with 4 R12000 processors (250 MHz) and 2GB RAM. The atom environment description method is combined with a structure editor and the trained neural networks in a new version of the software C_Shift (*27, 29*) that computes the chemical shift of organic substances. C_Shift is written in Microsoft Visual C++.

All ^{13}C chemical shifts obtained from the SpecInfo database were computed with the in-house version of SpecInfo (*14*) at the BASF AG in Ludwigshafen. ACD/CNMR (*16*), PredictIt NMR (*32*), Cosmos (*34*), Gaussian (*33*), SpecTool (*13*), and ChemDraw (*12*) were used to compute the chemical shifts of Taxol.

ACKNOWLEDGMENTS

We thank Dr. U. Sternberg (Cosmos GbR, Jena) for calculating the Taxol chemical shifts with the Cosmos 4.5 program package and M. Hollenbach (Bio-Rad Laboratories) for performing the chemical shift prediction for Taxol with PredictIt NMR 1.3. J. M. thanks Professor C. Griesinger for stimulating discussions and for providing computational resources in his laboratory at the University of Frankfurt, Germany. The authors thank B. Wedemeyer for useful discussions of the manuscript. J. M. thanks the Human Frontier Science Program (HFSP) for financial support.

REFERENCES

1. C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, and T. Wieland, MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation, *Anal. Chim. Acta* **314,** 141–147 (1995).
2. T. Lindel, J. Junker, and M. Köck, COCON: From NMR correlation data to molecular constitution, *J. Mol. Model.* **3,** 364–368 (1997).
3. J. Meiler and M. Will, Automated structure elucidation of organic molecules from 13C NMR spectra using genetic algorithms and neural networks, *J. Chem. Inf. Comput. Sci.* **41,** 1535–1546 (2001).
4. M. Schindler and W. Kutzelnigg, Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules, *J. Chem. Phys.* **76,** 1919–1933 (1982).
5. J. Gauss, Accurate calculation of NMR chemical shifts, *Beri. Bunsenges. Phys. Chem.* **99,** 1001–1008 (1995).
6. J. R. Cheeseman, G. W. Trucks, T. A. Keith, and M. J. Frisch, A comparison of models for calculating nuclear magnetic resonance shielding tensors, *J. Chem. Phys.* **104,** 5497 (1996).
7. J. R. Cheeseman and Æ. Frisch, Predicting Magnetic Properties with ChemDraw and Gaussian, *available at* http://www.gaussian.com/nmrcomp.pdf (2001).
8. D. M. Grant and E. G. Paul, Carbon-13 magnetic resonance. II. Chemical shift data for the alkanes, *J. Am. Chem. Soc.* **86,** 2984–2990 (1964).
9. L. P. Lindeman and J. Q. Adams, Carbon-13 nuclear magnetic resonance spectroscopy, *Anal. Chem.* **43,** 1245–1252 (1971).
10. J.-T. Clerc and H. Sommerauer, A minicomputer program based on additivity rules for the estimation of 13C NMR chemical shifts, *Anal. Chim. Acta* **95,** 33–40 (1977).
11. A. Fürst and E. Pretsch, A computer program for the prediction of 13C NMR chemical shifts of organic compounds, *Anal. Chim. Acta* **229,** 17–25 (1990).
12. Cambridge Soft Corporation, CS Chem Draw PRO 4.5 (1985–1997).
13. Upstream Solutions GMBH, NMR Prediction Products (SpecTool) (1998–2001).
14. BASF AG, SpecInfo (in-house version) (2002).
15. W. Bremser, HOSE—A novel substructure code, *Anal. Chim. Acta* **103,** 355–365 (1978).
16. Advanced Chemistry Development, ACD/CNMR 6.0 (1996–2001).
17. J. Zupan and J. Gasteiger, "Neural Networks for Chemists," VCH Verlagsgesellschaft mbH, Weinheim, 1993.
18. V. Kvasnicka, S. Sklenak, and J. Pospichal, Application of recurrent neural network in chemistry: Prediction and classification of 13C NMR chemical shifts in a series of monosubstituted benzenes, *J. Chem. Inf. Comput. Sci.* **32,** 742–747 (1992).
19. O. Ivanciuc, Artificial neural networks applications. Part 6. Use of nonbonded van der Waals and electrostatic intermolecular energies in the estimation of 13C-NMR chemical shifts in saturated hydrocarbons, *Rev. Roum. Chim.* **40,** 1093–1101 (1995).
20. D. Svozil, J. Pospichal, and V. Kvasnicka, Neural network prediction of carbon-13 NMR chemical shifts of alkanes, *J. Chem. Inf. Comput. Sci.* **35,** 924–928 (1995).
21. S. Thomas and E. Kleinpeter, Assignment of the 13C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network, *J. Prakt. Chem. Chem.-Ztg.* **337,** 504–507 (1995).

22. O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye, and J.-P. Doucet, 13C NMR chemical shift prediction of *sp*2 carbon atoms in acyclic alkenes using neural networks, *J. Chem. Inf. Comput. Sci.* **36,** 644–653 (1996).
23. Z. Li, Y. Huang, F. Hu, Q. Sheng, and S. Peng, Neural networks in spectroscopy: Estimation and prediction of chemical shifts of 13C NMR in alkanes by using subgraphs, *Bopuxue Zazhi* **14,** 507–514 (1997).
24. J. Meiler, R. Meusinger, and M. Will, Neural network prediction of 13C NMR chemical shifts of substituted benzenes, *Monats. Chem.* **130,** 1089–1095 (1999).
25. W. Robien, Das CSEARCH-NMR-Datenbanksystem, *Nachr. Chem. Tech. Lab.* **46,** 74–77 (1998).
26. C. Le Bret, A general 13C NMR spectrum predictor using data mining techniques, *SAR QSAR Env. Res.* **11,** 211–234 (2000).
27. J. Meiler, M. Will, and R. Meusinger, Fast determination of 13C-NMR chemical shifts using artificial neural networks, *J. Chem. Inf. Comput. Sci.* **40,** 1169–1176 (2000).
28. J. Meiler, E. Sanli, J. Junker, R. Meusinger, T. Lindel, M. Will, W. Maier, and M. Köck, Validation of structural proposals by substructure analysis and 13C NMR chemical shift prediction, *J. Chem. Inf. Comput. Sci.* **42,** 241–248 (2002).
29. J. Meiler, Software applications for chemistry, biochemistry and bioinformatics, *available at* www.jens-meiler.de (1996–2002).
30. P. B. Schiff, J. Fant, and S. B. Horwitz, Promotion of microtubule assembly *in vitro* by taxol, *Nature* **277,** 665–668 (1979).
31. H. J. Williams, A. I. Scott, R. A. Dieden, C. S. Swindell, and L. E. Chrilian, NMR and molecular modelling study of active and inactive taxol analogues in aqueous and nonaquous solution, *Can. J. Chem. Rev.* **72,** 252–260 (1994).
32. Bio-Rad, PreditcIt NMR 1.3 (2002).
33. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, Gaussian, Pittsburgh, PA, 2001.
34. U. Sternberg, F. Koch, and P. Losso, Cosmos 4.52 (2001).
35. U. Sternberg, Second sphere theory for the interpretation of chemical shifts, *Mol. Phys.* **63,** 249–267 (1988).
36. U. Sternberg and W. Priess, New semi-empirical approach for the calculation of 13C-chemical shift tensors, *J. Magn. Reson.* **125,** (1997).
37. W. Priess and U. Sternberg, Extreme fast calculation of 13C chemical shifts of large biomolecules, *THEOCHEM* **544,** 181–190 (2001).
38. D. Cherqaoi and D. Villemin, Use of a neural network to determine the boiling point of alkanes, *J. Chem. Soc. Faraday Trans.* **90,** 97–102 (1994).