

# Toward More Reliable $^{13}\text{C}$ and $^1\text{H}$ Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches

Yegor D. Smurnyy,<sup>†</sup> Kirill A. Blinov,<sup>†</sup> Tatiana S. Churanova,<sup>†</sup> Mikhail E. Elyashberg,<sup>†</sup> and Antony J. Williams<sup>\*,‡,§</sup>

Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation, and Advanced Chemistry Development, Inc., 110 Yonge Street, 14th Floor, Toronto, Ontario, Canada M5C 1T4

Received July 17, 2007

The efficacy of neural network (NN) and partial least-squares (PLS) methods is compared for the prediction of NMR chemical shifts for both  $^1\text{H}$  and  $^{13}\text{C}$  nuclei using very large databases containing millions of chemical shifts. The chemical structure description scheme used in this work is based on individual atoms rather than functional groups. The performances of each of the methods were optimized in a systematic manner described in this work. Both of the methods, least-squares and neural network analyses, produce results of a very similar quality, but the least-squares algorithm is approximately 2–3 times faster.

## 1. INTRODUCTION

The accurate and robust prediction of NMR chemical shifts has both practical and theoretical interest. The *de novo* structure elucidation of natural products, the verification of chemical structures contained within synthetic libraries, and the potential for enhancing the experience of chemical education are just a few areas which could derive value from NMR shift prediction. NMR prediction is now a tool utilized not only by NMR spectroscopists but also by synthetic chemists. Contemporary robotic open-access spectrometers provide convenient access to both 1D and 2D NMR spectra for chemists to assist in structure verification and elucidation.  $^1\text{H}$  and  $^{13}\text{C}$  1D spectra are the primary tools utilized by chemists for structure verification, and the provision of prediction tools provides reasonable to excellent accuracy in the quality of NMR prediction (*vide infra*) and thereby provides a means to speed up the analysis of spectra using a strong basis of prior knowledge. In recent years, the utilization of NMR chemical shift prediction as an integral part of an expert system intended for computer-aided structure elucidation (CASE) has been common. For this purpose, both  $^{13}\text{C}$  and  $^1\text{H}$  NMR spectral predictions are used for the identification of the most probable structure (see review in ref 1 and refs 2 and 3).

Historically, two main classes of algorithms have been developed: database-based and rules-based prediction algorithms. For the first approach, a large database of chemical structures with associated chemical shifts is compiled. For each structure, a descriptor is assigned which reflects its major structural features. Then, when the database is queried with the descriptor, similar structures are identified, and the resulting values are weighted averages of the experimental data corresponding to these structures. A number of commercial databases and associated prediction software are

currently available including CSEARCH,<sup>4,5</sup> ACD/Labs,<sup>6</sup> and SpecInfo<sup>7</sup> as well as a publicly available database known as NMRShiftDB.<sup>8</sup> The most popular structure description algorithm is the Hierarchical Organization of Spherical Environments (HOSE) code,<sup>9</sup> despite its supposed poor performance with unusual structures and the slow speed of prediction due to relatively slow database engines.

Another class of algorithm utilizes a database to extract a set of additivity rules which allow for the rapid calculation of chemical shift values for atoms of interest. In an ideal case, these algorithms would perform similarly upon a diverse set of chemical structures. This approach was originally suggested in 1964 in the pioneering work of Grant and Paul<sup>10</sup> for the prediction of the  $^{13}\text{C}$  chemical shifts of alkanes. The method was then extended to many different classes of organic compounds and then also applied to  $^1\text{H}$  chemical shift prediction: Reed presented a scheme for the prediction of chemical shifts in substituted benzenes.<sup>11</sup> The prediction rules published in the literature were later generalized in the monograph of Pretsch et al.<sup>12</sup> The first computer program for  $^{13}\text{C}$  chemical shift prediction on the basis of additivity rules was presented by Clerc and Sommerauer.<sup>13</sup> Later, Pretsch and co-workers developed programs to provide fast prediction of both  $^{13}\text{C}$  and  $^1\text{H}$  chemical shifts.<sup>14–18</sup> Traditionally, least-squares regression techniques have been used<sup>19,20</sup> to formulate the rules. However, since the 1990s, there have been increasing efforts by researchers to use neural networks as an alternative and potentially more flexible and powerful approach.<sup>21–23</sup>

Despite decades of intensive research, there are a number of questions not yet fully addressed in the field of neural network applications. First, very few research groups (for example, ref 24) have systematically compared available approaches such as partial least-squares (PLS) versus neural networks (NN). Second, many studies<sup>23,25–27</sup> focus on a limited class of compounds. This approach biases the results due to the lack of structural variability but severely limits practical applications. Third, we believe that it is necessary to separate the two parts of an algorithm: (a) encoding a

\* Corresponding author e-mail: antony.williams@chemspider.com.

<sup>†</sup> Advanced Chemistry Development, Moscow Department.

<sup>‡</sup> Advanced Chemistry Development, Toronto.

<sup>§</sup> Current address: ChemZoo Inc., 904 Tamaras Circle, Wake Forest, NC-27587.

chemical structure into a numerical input, the number and nature of the descriptors, the involvement of stereochemistry, and so forth, and (b) the regression process, either by least-square methods or neural network methods. These two parts should be evaluated and compared separately as is discussed in this work but, to the best of our knowledge, absent in earlier works.

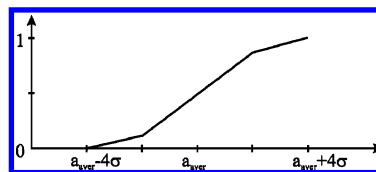
In this article, we seek to address each of the above-mentioned issues using  $^{13}\text{C}$  chemical shifts as the data for analysis. We initially optimized the description scheme used throughout our studies; then, the parameters affecting the performance of the neural networks were tuned to achieve optimal performance. We then compared both PLS and neural networks using one of the largest available databases<sup>6</sup> for the purpose of training. Finally, we applied our findings from our studies for  $^{13}\text{C}$  chemical shifts to examine the performance of  $^1\text{H}$  chemical shift prediction to demonstrate the general applicability of our findings.

## 2. METHODS

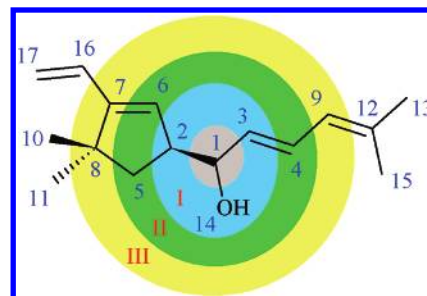
**2.1. Databases.** The carbon and proton chemical shift databases used in this work are comprised of approximately 2 million  $^{13}\text{C}$  and 1.2 million  $^1\text{H}$  chemical shift values, respectively. Care was taken to avoid overlap between the data sets used for NN training and comparison. The training data set was compiled using experimental data published from the early 1990s until 2004. Spectral assignments were taken from articles published in a number of journals, including *J. Org. Chem.*, *Tetrahedron*, *J. Nat. Prod.*, and so forth. During each run, 3% of the data set was randomly chosen to be used as a test set during training. A comparison of the algorithms reported in this article was performed using a completely independent database compiled from data originally published in 2005 and consisting of approximately 118 000  $^{13}\text{C}$  and 116 000  $^1\text{H}$  chemical shifts.

**2.2. Algorithms.** The effectiveness of different neural network configurations was tested using custom scripts written in the MATLAB programming language (version 6.0.; The MathWorks, Natick, MA) and using the associated Neural Networks toolbox. The neural network and partial least-square regression algorithms were then coded using the Borland Delphi 5.0 environment, Object Pascal, and assembler programming languages. The LAPACK library<sup>28</sup> was used for matrix computations. The detailed mathematical descriptions of neural network<sup>29</sup> and nonlinear regression<sup>30</sup> analyses can be found elsewhere. All programs were executed on Intel PC workstations running under the Windows XP operating system using CPUs operating with clock speeds of 1.8–3.2 GHz and 1–4 GB of memory installed.

All of the neural networks tested in this work include an input layer, one to four hidden layers, and a single output neuron. Logistic or hyperbolic tangent activation functions were used. Networks were trained using the standard back-propagation algorithm.<sup>29</sup> Weights were updated after processing each pattern with both the training speed and the momentum set to 0.5. Neural network input and output values were assumed to be within the interval of [0;1]. The tri-linear function shown in Figure 1 was used to map an arbitrary interval onto [0;1]. This choice is due to the simplicity and robustness of the function. Simultaneously,



**Figure 1.** Tri-linear function used to map the input values onto the [0;1] interval utilized by our neural networks.  $a_{\text{aver}}$  is the mean value of the  $a$  variable, and  $\sigma$  is the standard deviation.



**Figure 2.** Encoding of the atom environment. The three nearest spheres are marked as cyan, green, and yellow circles, and the Roman numerals denoting the number of each sphere are shown. Atom no. 1 is the central atom. The blue numbers are assigned arbitrarily and serve as references (see the text for more details).

this tri-linear function allows the algorithm to cover a wider interval than a simple linear function. Such a tri-linear scaling scheme allows the prediction of chemical shift values up to four standard deviations away from the mean value, which, assuming a Gaussian distribution of values, accounts for 99.9% of the data.

For the network inputs  $x^i$ , the corresponding values of an average of  $a_{\text{aver}} = x_{\text{aver}}^i$ , and the associated standard deviations were calculated using all available training patterns in which the input has a nonzero value. The values for the chemical shifts were calculated using all available training patterns subsequently used for reverse transformation from the [0;1] output to a chemical shift value in parts per million delivered to the user.

**2.3. Data Encoding.** To encode a chemical structure into a numerical representation, an atom-based scheme was used as shown in Figure 2.

The environment surrounding an atom is divided into spheres, each including all atoms and separated from the center by a definite number of covalent bonds. In our work, we typically account for the nearest six spheres or less. Every atom within a sphere is classified into one of the predefined atomic classes described in Table 1. The scheme was inspired by earlier works in this field.<sup>22,24</sup> We have added extra features such as a more complete list of nontypical atoms and the ability to take solvents into account in the prediction algorithm.

Some additional parameters improve chemical shift prediction for atoms included in conjugated systems. For the purpose of this work, we define a conjugated system as atoms forming conjugated double bonds plus all of their immediate neighbors. If a central atom participates in a system, then all other atoms in the system are marked with a special flag. For example, in the molecule shown in Figure 2, the central atom marked as number 1 is part of a conjugated system. As a result, atoms 3, 4, 9, and 12 (the diene system) and their neighbors 1 (the center), 13, and 15 will be marked. Note that atoms 6, 7, 16, and 17 are not marked since they

**Table 1.** Atomic Classes Used to Classify Atoms<sup>a</sup>

carbon	sp <sup>3</sup> (C, CH, CH <sub>2</sub> , CH <sub>3</sub> ), sp <sup>2</sup> (C, CH, CH <sub>2</sub> ), sp (C, =C=), aromatic (C, CH), carbonyl
heteroatoms	(3-)N, (2-)NH, NH <sub>2</sub> , =NH, N(sp), N(V), aromatic N, (2-)O, OH, =O, (3-)P, P(V), (2-)S, =S, S(IV), S(VI), F, Cl, Br, I
exotic elements	Si, Ge, Sn, (2-)Se, =Se, (2-)Te, =Te, B, As(III), As(V), Sb(III)
solvent	CHCl <sub>3</sub> , CH <sub>2</sub> Cl <sub>2</sub> , C <sub>6</sub> H <sub>6</sub> , (CH <sub>3</sub> ) <sub>2</sub> SO, dioxane, CH <sub>3</sub> OH, CH <sub>3</sub> NO <sub>2</sub> , tetrahydrofuran, cyclohexane, (CH <sub>3</sub> ) <sub>2</sub> CO, CH <sub>3</sub> CN, (CH <sub>3</sub> ) <sub>2</sub> CONH <sub>2</sub> , pyridine, CF <sub>3</sub> COOH, CH <sub>3</sub> COOH, C <sub>6</sub> H <sub>5</sub> NO <sub>2</sub> , C <sub>6</sub> H <sub>5</sub> CH <sub>3</sub> , CCl <sub>4</sub> , CS <sub>2</sub> , H <sub>2</sub> O, other/unknown
other parameters	formal positive charge, formal negative charge, total count of hydrogen atoms, total count of ring closures, involvement into the same $\pi$ -conjugated system with the central atom, Z or E double bond, stereochemistry

<sup>a</sup> The adjective "aromatic" means that the atom is within an aromatic system. The symbol "(*n*-)X" designates that *n* single bonds are attached to the X atom.

are separated from the central atom by two  $\sigma$  bonds. Additional flags were also used to take into account double-bond stereochemistry. If we use atom 4 as the central atom, then atoms 13 and 15 both lie in the third sphere and both have equivalent descriptors. The addition of stereo descriptors allows these atoms to be distinguished. Atom 15 lies on the same side as the atom of the double-bond marked by 9–12, while atom 13 lies on the opposite side. Atom 15 is therefore marked as the Z atom, while atom 13 is marked as the E atom.

Stereochemical descriptors are not implemented systematically throughout the system. However, a separate flag is set for atoms which are one bond away from a three- to six-membered aliphatic ring. These atoms are classified as located on either the same or opposite side of a ring. In the molecule shown in Figure 2, atom 10 lies on the same side of the five-membered ring made up of atoms 2, 6, 7, 8, and 5 while atom 11 lies on the opposite side relative to the central atom 1. This method obviously can only be used for relatively rigid rings and is inapplicable to the stereochemistry of flexible systems such as large rings and chains.

In many cases, so-called "cross-increments" were used. These refer to pairs of atoms—for each two atoms separated by a small number of bonds, an independent identifier is generated and stored. In this study, we considered pairs separated by one to three bonds, or by one to four bonds in conjugated systems, with both atoms located within the first three to four spheres. For example, if we use atom 1 as the central atom, then the following pairs of atoms, 3–14, 3–2, and 2–14 (in the first sphere) and 3–4, 2–6, 2–5, 3–6, 3–5, 2–4, 5–6, 4–14, 5–14, and 6–14 in the second sphere, should be taken into account. Atom pairs 4–5 and 4–6 are ignored since the distance between the atoms is too long in this case—four bonds. The distances between atom and sphere numbers are also used to describe cross-increments. For example, the atoms in pair 3–4 are both CH (sp<sup>2</sup> hybridized) and lie in the first and second spheres, correspondingly, and are separated by one bond. The atoms in pair 3–6 are also both CH (sp<sup>2</sup> hybridized), lie in the first and second spheres, and are separated by three bonds.

**Table 2.** Results of Some of the Test Runs Using a MATLAB Implementation of the Neural Network Algorithm<sup>a</sup>

no.	number of neurons in hidden layers	transfer function	cross- increments	mean error, ppm (test data set)
1	20–10	logistic	two bonds	3.06
2	30	logistic	two bonds	3.07
3	50	logistic	two bonds	3.12
4	20–10	tanh	two bonds	3.15
5	30	logistic	one bond	3.23
6	20–10	logistic	one bond	3.23
12	20–10	logistic	none	3.68
18	50	tanh	none	4.58

<sup>a</sup> The mean error of the test set is shown as a function of the neural network configuration. The data are sorted by the mean error in ascending order, with some entries being omitted for clarity.

### 3. RESULTS AND DISCUSSION

#### 3.1. Optimization of the Neural Network Performance.

Since the performance of neural network algorithms depends on a large set of factors, we first performed a set of test calculations to try and establish an optimal parameter set using a smaller test data set. For this purpose, we randomly selected 32 000 structures from the main training database. We also questioned whether a neural network would benefit from the inclusion of cross-increments. One could argue that, since neurons in hidden layers receive inputs from all input neurons, weights can be adjusted to take into account the simultaneous occurrence of specific atomic types. For example, the carbonyl and the hydroxyl groups together would represent a carboxylic group and would render cross-increments unnecessary.

Using a MATLAB implementation of the neural network algorithms, we set up 18 test runs while varying a few parameters in order to investigate the variation in performance as a result of changes in the different parameters.

- The number of cross-increments varied as (1) none, (2) pairs of atoms separated only by one bond, or (3) pairs of atoms separated by up to two bonds.

- The geometry of the neural networks was (1) 50 hidden neurons in one layer, (2) 30 hidden neurons in one layer, or (3) two hidden layers of 20 and 10 neurons.

- The transfer function was either the logistic function  $f(x) = (1 + e^{-x})^{-1}$  or hyperbolic tangent (tanh).

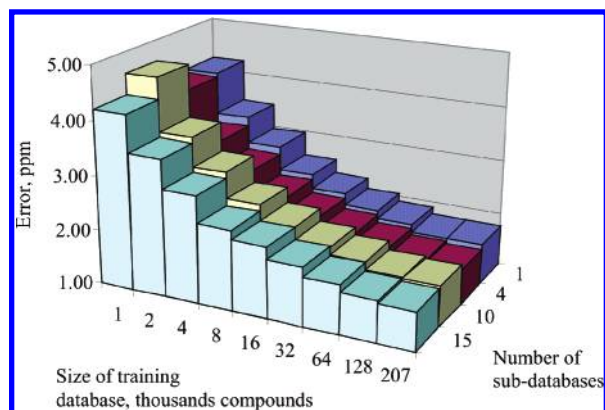
As shown in Table 2, the top configurations 1–4 took into account the maximum available number of cross-increments. The best result without cross-increments is shown for comparison and is only ranked 12th out of the 18 runs. At the same time, it is clear that arrangement of the neurons into layers plays a somewhat secondary role since the difference between the first two networks with the same total number of neurons but with different geometries is negligible.

These results define a reference point for further studies. While subsequently scaling up our calculations to approximately 2 000 000 <sup>13</sup>C chemical shifts, we mainly utilized networks with one to three hidden layers and with a logistic transfer function. We have also tried to provide as many cross-increments as possible to our networks with only time and memory restrictions limiting our analyses.

#### 3.2. Choosing the Optimal Number of Subdatabases.

While dealing with a large database is likely to be of value for analysis purposes, we chose to split it into a number of smaller subsets. The most popular strategy for splitting a





**Figure 3.** Dependence of the mean error derived from the test set for a number of subdatabases relative to the size of the training set. The results were obtained with neural networks containing 30 neurons in one hidden layer. The splitting schemes leading to 1, 4, 10, and 15 different classes are described in the text.

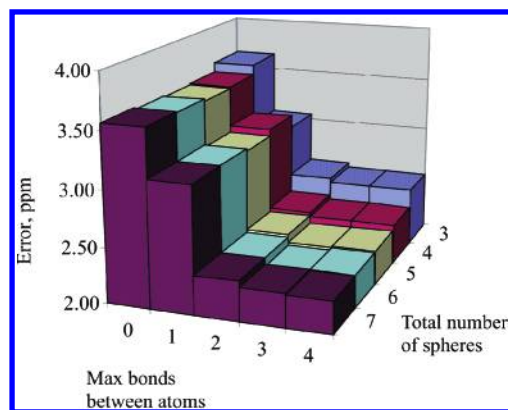
large database is based on the nature of the central atom type. All carbons can be classified according to hybridization, the number of attached protons, or both. Also, some groups are so specific and abundant, for example, a carbonyl group, that they might deserve a separate class. We have utilized four different strategies:

- The database was used as a whole.
- The carbon atoms were classified only according to hybridization, three aliphatic classes ( $\text{sp}^3$ ,  $\text{sp}^2$  and  $\text{sp}$ ) and one for all carbons within aromatic rings, to give a total of four different classes.
- The hybridization and the number of attached protons were both taken into account. This led to 10 classifications: four  $\text{sp}^3$  (C, CH,  $\text{CH}_2$ , and  $\text{CH}_3$ ), three  $\text{sp}^2$  (C, CH, and  $\text{CH}_2$ ), one  $\text{sp}$ , and two aromatic (C and CH).

The most detailed scheme includes states of hybridization, the number of protons attached to the carbon atom, and the presence or absence of a heteroatom within one bond of a central atom. The 15 resulting classes include aliphatic  $\text{sp}^3$  (seven classes: C, C(het), CH, CH(het),  $\text{CH}_2$ ,  $\text{CH}_2$ (het), and  $\text{CH}_3$ ), aliphatic  $\text{sp}^2$  (four classes: C, CH,  $\text{CH}_2$ , and CO), aliphatic  $\text{sp}$ , and three aromatic classes (C, C(het), and CH). The symbol (het) denotes a heteroatom, oxygen, or nitrogen, in this case, nearby. Some classes were merged into one subclass, for example, aliphatic  $\text{sp}^3$   $\text{CH}_3$  and  $\text{CH}_3$ (het), because a smaller class was too small or not diverse enough to reliably teach a network, especially one with a higher number of neurons.

It appears that more detailed classifications bring more flexibility. For example, one can introduce specific description schemes for certain atomic types—chemical shifts of the methyl group might highly depend on stereochemistry, and it is possible to introduce stereo descriptors specifically for this atomic type. Smaller databases are also easier to handle in terms of computer memory requirements. However, restricting the training set to very similar compounds prevents a neural network from making generalizations and decreases the quality of the results.

As seen in Figure 3, the differences between classifications are more apparent with a smaller number of training pools. These results suggest that, for a large enough database, about 50 000–100 000 compounds, the results are only slightly dependent on the classification scheme. For the largest



**Figure 4.** Mean error of a set as a function of the maximum number of bonds between the atoms in the cross-increments and the number of spheres used to describe an atom's neighborhood.

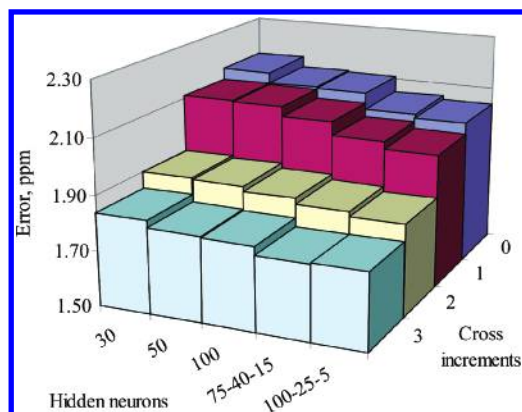
database utilized here, over 207 000 compounds, the difference between the best classification, four centers giving an error of 1.65 ppm, and the worst, one center giving an error of 1.75 ppm, was only 0.10 ppm. This difference is not practically important for carbon NMR shifts. We decided to utilize the scheme with 15 classes for commercial purposes due to the flexibility discussed earlier.

**3.3. Choosing the Optimal Number of Cross-Increments.** Using cross-increments is one of the few ways to boost the performance of a regression-based scheme. Although the approach is less popular with neural networks, it was illustrated above that the latter can also benefit from explicitly specified cross-increments. At the same time, providing too many increments and cross-increments leads to an unnecessarily detailed description of a structure and, consequently, to overtraining. Using a small part of the whole database, 16 000 compounds with 212 000 chemical shifts, we systematically changed a few parameters to elucidate an optimal scheme. Since it is possible to vary the number of atomic spheres under consideration, we tried values from 3 to 7. Due to limited computational resources, it is impossible to generate all possible cross-increments between atoms up to the sixth and seventh spheres. In this study, cross-increments were created for atoms located no further than the third sphere. We varied the maximum number of bonds between the atoms in the cross-increments from zero (no cross-increments) to four.

For these tests, our main training database was split into 10 subdatabases (see above), and a PLS routine was used.

Figure 4 summarizes our findings. The best result was obtained from a combination of six spheres and three bonds to provide a mean error of 2.27 ppm. There is practically no difference between the two largest sets of increments with up to three or four bonds between the atoms, and this suggests that further refinement might not be necessary and can lead to overfitting. The same applies to the total number of spheres—six are enough. This makes perfect chemical sense since there are few electrostatic interactions or conjugation, inductive, or mesomeric effects in aromatic systems that span across more than five or six covalent bonds.

**3.4. Comparison of  $^{13}\text{C}$  Chemical Shift Prediction by Neural Networks and Partial Least Squares.** After optimization of a few key features such as the structure description algorithm and the NN parameters, we compared the PLS and NN methods. For the neural networks, we varied



**Figure 5.** Mean  $^{13}\text{C}$  chemical shift error as a function of the neural network geometry and the maximum number of bonds separating atoms in the cross-increments. See text for details.

the size of the network, varying the number of hidden layers, the quantity of neurons, and the number of cross-increments provided. All other parameters such as the transfer function, the number of subdatabases, and so forth were taken as optimized during the previous steps. An independent test set of 118 000 individual  $^{13}\text{C}$  chemical shifts was used for comparison. Calculations were performed with 15 individual subdatabases as described above. All cross-increments were constructed from atoms separated by not more than one covalent bond and located within one to three spheres from the central atom.

As shown in Figure 5, the performance of neural networks gradually increases upon adding more hidden neurons and cross-increments. Not surprisingly, the best result, providing a mean error of 1.77 ppm, was achieved with the maximum number of cross increments and the largest neural network. It should be noted that, for the various network configurations compared, the total number of parameters for each network is not the same and all networks with more than one layer have more parameters. That said, the only extra parameter, and therefore additional degree of freedom, is the number of neurons in the additional hidden layers. This additional single extra parameter was optimized by comparing the two configurations of 100–25–5 and 75–40–15, both with the same total number of neurons. In general, it is concluded that the total number of neurons matters but not their arrangement.

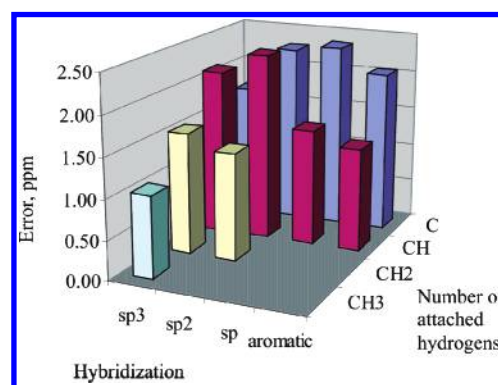
The configuration was tuned by adding more cross-increments up to the third sphere with no more than two covalent bonds between the atoms and using stereo information in the atomic descriptors (see above). With three layers of 100, 25, and 5 hidden neurons, the network produced a mean error of 1.59 ppm with a largest error value of 85 ppm and 0.6% of the chemical shifts predicted with an error of more than 10 ppm (Table 3).

A similar approach was used to evaluate the performance of the PLS algorithm. The optimal number of latent variables was determined using an independent test data set. Calculations were performed on this control data set using up to 200 latent variables, and the number providing the lowest average deviation was used for further analysis. The optimal number of latent variables is different for different types of carbon environments and varies between 50 and 150. The same database split of 15 subdatabases was used. For the cross-increments, we used a configuration of six spheres,

**Table 3.** Best Results for the Prediction of Both  $^{13}\text{C}$  and  $^1\text{H}$  Chemical Shifts by HOSE<sup>6</sup> Codes, PLS, and NN Algorithms<sup>a</sup>

type of spectrum	prediction method	mean error, ppm	standard deviation, ppm	maximum error, ppm	percent of outliers
$^{13}\text{C}$	HOSE	1.81	3.05	57.99	2.8%
	PLS	1.71	2.61	51.57	0.7%
	NN	1.59	2.45	85.82	0.6%
$^1\text{H}$	HOSE	0.19	0.30	3.94	1.3%
	PLS	0.18	0.26	2.72	0.7%
	NN	0.18	0.26	3.71	0.8%

<sup>a</sup> The specific set of parameters for each method is described in the text. Outliers are defined as those shifts predicted with an error of more than 10 ppm for  $^{13}\text{C}$  predictions and more than 1.0 ppm for  $^1\text{H}$  predictions.



**Figure 6.** Dependence of the mean error (ppm) on the number of substituents and the hybridization of the central atom. The data shown are for  $^{13}\text{C}$  chemical shift predictions with the most efficient neural network configurations (see text for details).

and cross-increments were taken from three first spheres with atoms separated by not more than three covalent bonds. This setup was shown to be the most effective in our previous studies (see above). The results obtained were of a similar quality. The mean  $^{13}\text{C}$  chemical shift error is slightly higher (1.71 ppm), but other benchmark parameters were close to those obtained for the neural network (Table 3). The number of severe outliers is somewhat lower with only 0.7% of the centers producing errors of greater than 10 ppm. The maximum error was 56 ppm.

One may request an explanation for why the neural-network-based  $^{13}\text{C}$  chemical shift calculations give the best accuracy and yet have the largest maximum error. This is due to the fact that the ability to predict values for the most representative compounds in a given class while also providing optimal performance for the most exotic compounds is a significant challenge. Predictions can be optimized either for “typical” or “atypical” members of a data set, and as a result, we observe good prediction performance for “average structures” but poor performance for the more exotic compounds.

Values for the mean error determined using NN and presented in Figure 6 as a function of the number of substituents (i.e.,  $\text{CH}_2$  vs  $\text{CH}$  for the same hybridization state) suggest that generally the errors are higher for more highly substituted atoms. This would appear reasonable because the cross-influence between the different substituents enhances the nonlinear effects.

The comparison shows that both methods can provide results of similar quality after being properly optimized. A

**Table 4.** The General Workflow of the Analysis<sup>a</sup>

step	nucleus	algorithm	size of the database, chemical shifts		size of the database, compounds	
			training	testing	training	testing
optimization of neural network parameters	$^{13}\text{C}$	NN	422 000	12 660	29 000	3000
optimization of number of subdatabases	$^{13}\text{C}$	PLS	2 000 000	118 000	207 000	11 000
optimization of number of structure descriptors	$^{13}\text{C}$	PLS	212 000	118 000	16 000	11 000
comparison of least-square regression vs neural networks	$^{13}\text{C}$	PLS, NN	2 000 000	118 000	207 000	11 000
	$^1\text{H}$	PLS, NN	1 150 000	116 000	189 000	14 000

<sup>a</sup> The sizes of the databases are represented by the number of chemical shifts as well as the number of compounds.

neural network, in general, seems to perform better with atoms whose chemical shifts are closer to an average value for the corresponding atomic type. Linear regression can more easily handle exotic fragments (such as in the compound  $\text{Cl}_4$  with a chemical shift of  $-292.5$  ppm), since even the most unusual combination of substituents can easily be assigned with an appropriate incremental value leading to a more accurate prediction. At the same time, values of weights for most of the neurons in a network are affected by all of the structures present during the training process, and the impact of unusual structures is masked by the majority of the more regular structures. These differences are highlighted by the lower value of the mean error for the neural network. Simultaneously, the maximum error being more sensitive to a small number of badly predicted structures is better for the incremental scheme.

Both of these methods perform better than our implementation<sup>6</sup> of the database-based HOSE code approach<sup>9</sup> (Table 3). For a given atom, the algorithm<sup>6</sup> retrieves few structures from a database which have chemically similar nuclei. The predicted value is the weighted average of chemical shifts contained within the database structures. The approach fails with structures which are underrepresented in the database, and this raises the standard deviation and the maximum error.

**3.5. Prediction of  $^1\text{H}$  Chemical Shifts.** An attempt was made to apply the experiences from the  $^{13}\text{C}$ -related analyses to the prediction of proton chemical shifts. Over 1 million chemical shifts were used as a training set, and the same test data set was that used for the  $^{13}\text{C}$  analysis, with a total of 114 494  $^1\text{H}$  chemical shifts.

The whole training database was split into nine sub-datasets, namely, the aliphatic  $\text{sp}^3$  (five classes:  $\text{CH}$ ,  $\text{CH}(\text{het})$ ,  $\text{CH}_2$ ,  $\text{CH}_2(\text{het})$ , and  $\text{CH}_3$ ), aliphatic  $\text{sp}^2$ , aromatic  $\text{sp}^2$ , aliphatic  $\text{sp}$ , and protons attached to heteroatoms. The notation (het) denotes a group attached to a heteroatom. The neighborhood of an atom was described in exactly the same manner as described earlier for the  $^{13}\text{C}$  studies. All additional flags specified in the section above (stereo configuration, Z/E conjugated system, etc) were included.

The neural network used for the calculation had 100, 25, and 5 hidden neurons arranged in three layers. The six nearest spheres were used; cross-increments were constructed from atoms in the first three layers separated by not more than one covalent bond. PLS regression was performed within the same six-sphere vicinity, though with more cross-increments: within the three nearest spheres with atoms separated by three covalent bonds or less.

The best results for the  $^1\text{H}$  chemical shifts predictions are presented in Table 3. The neural networks and PLS approaches performed in a remarkably similar manner. The quality of results is much less dependent on the number of cross-increments—most of the NN and PLS configurations result in a mean error of approximately 0.2 ppm. In order to reduce the error to that experienced in experimental determinations, further optimization such as the detailed description of the 3D geometry might be necessary. Table 4 summarizes the separate steps carried out in this work.

**3.6. Speed of Chemical Shift Prediction.** Modern CASE expert systems are based on the utilization of 2D NMR data, and this allows the identification of isolated materials such as natural products or synthesized organic molecules. These systems are capable of elucidating large molecules containing 100 or more skeletal atoms.<sup>1</sup> Since the initial structural information extracted from 2D NMR data is fuzzy by nature,<sup>1</sup> the number of structures that are consistent with the spectral data can usually be rather large (up to tens of thousands<sup>3</sup>). As a result, the selection of the most probable structure from a large output file containing many molecules requires an approach whereby the expert systems can utilize both *accurate* and *fast* approaches for NMR chemical shift prediction.

Both the  $^{13}\text{C}$  and  $^1\text{H}$  chemical shift calculations compared in this study were implemented into a software expert system known as *Structure Elucidator*.<sup>2,3</sup> The prediction speed was estimated by the spectral prediction of candidate structures generated by the program. The average speed of the  $^{13}\text{C}$  chemical shift prediction by PLS is about 9000 shifts per second on a 2.8 GHz computer, while the neural-network-based algorithm was approximately 2.5–3 times slower. The combination of this high speed of prediction with an appropriate accuracy for prediction with an average deviation of 1.71 ppm makes the PLS approach a powerful tool for computer-aided structure elucidation. A study regarding the impact of these new NMR spectrum prediction tools in regard to the possibility of improving CASE systems is in progress, and the results will be reported in a future publication.

#### 4. CONCLUSIONS

The results of both  $^{13}\text{C}$  and  $^1\text{H}$  NMR chemical shift predictions using neural networks and partial least-squares algorithms have been systematically compared. Two of the largest databases currently available were utilized in this work: over 2 million  $^{13}\text{C}$  shifts and 1 million  $^1\text{H}$  shifts. In



most cases, these two algorithms performed remarkably similarly. It is concluded that neural networks do not automatically lead to better results and that a simpler least-squares approach might still be the method of choice. It was determined that the way a chemical structure was converted into a numerical representation provided the most influence regarding the results obtained. Both methods required rather detailed descriptions of the structure and included atoms up to six covalent bonds apart. It has been shown that neural networks still require cross-increments and are unable to recognize even the most common functional groups such as a carbonyl group from a description based solely on the individual atoms.

The quality of results obtained using our current setup is probably limited to average errors of 1.5 ppm for carbon and 0.2 ppm for protons. Our belief is that any further improvement in the precision of NMR prediction will be due to the perfection of the structure description algorithms and, more importantly, the quality of training databases rather than the method of regression used to extract the prediction algorithms. An important result from this study is that the incremental approach now implemented in ACD/CNMR Predictor provides an average speed for  $^{13}\text{C}$  chemical shift prediction of over 9000 shifts per second with a mean chemical shift deviation of 1.7 ppm. This is a significant improvement in both speed and accuracy over previously available algorithms.

## REFERENCES AND NOTES

- (1) Elyashberg, M. E.; Williams, A. J.; Martin, G. E. Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. *Prog. NMR Spectrosc.* **2007**, in press.
- (2) Elyashberg, M. E.; Blinov, K. A.; Molodtsov, G. S.; Williams, A. J.; Martin, G. E. Structure Elucidator: A Versatile Expert System for Molecular Structure Elucidation from 1D and 2D NMR Data and Molecular Fragments. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 771–792.
- (3) Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, G. S.; Martin, G. E. Are deterministic expert systems for computer assisted structure elucidation obsolete? *J. Chem. Inf. Model.* **2006**, *46*, 1643–1656.
- (4) Robien, W. CSEARCH. <http://nmrpredict.orc.univie.ac.at/> (accessed June 25, 2007).
- (5) Robien, W. Das CSEARCH-NMR-Datenbanksystem. *Nachr. Chem., Tech. Lab.* **1998**, *46*, 74–77.
- (6) ACD/NMR Predictors; Advanced Chemistry Development: Toronto, Canada, 2007.
- (7) Bremser, W.; Grzonka, M. SpecInfo - A multidimensional spectroscopic interpretation system. *Microchim. Acta* **1991**, *104*, 1–6.
- (8) Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB - Constructing a Free Chemical Information System with Open-Source Components. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733–1739.
- (9) Bremser, W. HOSE - A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *2*, 355–365.
- (10) 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.
- (11) Reed, J. J. R. Structure Determination of Substituted Benzenes by Proton Magnetic Resonance. Empirical Treatment of Substituent Effects and Their Utility in Predicting Chemical Shifts. *Anal. Chem.* **1967**, *39*, (13), 1586–1593.
- (12) Pretsch, E.; Clerc, T. E.; Seibl, J.; Simon, W. *Tables of Spectral Data for Structure Determination of Organic Compounds*; Springer-Verlag: Berlin, 1989.
- (13) Clerc, J.-T.; Sommerauer, H. A Minicomputer Program Based on Additivity Rules for the Estimation of Carbon-13 NMR Chemical Shifts. *Anal. Chim. Acta* **1977**, *95*, 33–40.
- (14) Fürst, A.; Pretsch, E. A Computer Program for the Prediction of Carbon-13 NMR Chemical Shifts of Organic Compounds. *Anal. Chim. Acta* **1990**, *229*, 17–25.
- (15) Pretsch, E.; Fürst, A.; Badertscher, M.; Burgin, R.; Munk, M. E. C13shift - a Computer-Program for the Prediction of Carbon-13 NMR-Spectra Based on an Open Set of Additivity Rules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 291–295.
- (16) Schaller, R. B.; Arnold, C.; Pretsch, E. New Parameters for Predicting  $^1\text{H}$  NMR Chemical Shifts of Protons Attached to Carbon Atoms. *Anal. Chim. Acta* **1995**, *312*, 95–105.
- (17) Schaller, R. B.; Pretsch, E. A Computer Program for the Automatic Estimation of  $^1\text{H}$  NMR Chemical Shifts. *Anal. Chim. Acta* **1994**, *290*, 295–302.
- (18) Upstream Solutions GMBH, NMR Prediction Products (SpecTool). <http://www.upstream.ch/products/nmr.html> (accessed May 28, 2007).
- (19) Clouser, D. L.; Jurs, P. C. Simulation of the  $^{13}\text{C}$  Nuclear Magnetic Resonance Spectra of Ribonucleosides Using Multiple Linear Regression Analysis and Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168–172.
- (20) Lucic, B.; Amic, D.; Trinajstić, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks on Three QSPR Data Sets. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403–413.
- (21) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P.  $^{13}\text{C}$  NMR Chemical Shift Prediction of  $\text{sp}^2$  Carbon Atoms in Acyclic Alkenes Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644–653.
- (22) Meiler, J.; Meusinger, R.; Will, M. Fast Determination of Carbon-13 NMR Chemical Shifts Using Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169–1176.
- (23) Kaur, J.; Brar, A. S. An Approach to Predict the C-13 NMR Chemical Shifts of Acrylonitrile Copolymers Using Artificial Neural Network. *Eur. Polym. J.* **2007**, *43*, 156–163.
- (24) Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using Neural Networks for Carbon-13 NMR Chemical Shift Prediction-Comparison with Traditional Methods. *J. Magn. Reson.* **2002**, *157*, 242–252.
- (25) Bosque, R.; Sales, J. A QSPR Study of the  $^{31}\text{P}$  NMR Chemical Shifts of Phosphines. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 225–232.
- (26) Da Costa, F. B.; Binev, Y.; Gasteiger, J.; Aires-De-Sousa, J. Structure-based predictions of  $^1\text{H}$ -1 NMR Chemical Shifts of Sesquiterpene Lactones Using Neural Networks. *Tetrahedron Lett.* **2004**, *45*, 6931–6935.
- (27) Beger, R. D.; Harris, S.; Xie, Q. Models of Steroid Binding Based on the Minimum Deviation of Structurally Assigned  $^{13}\text{C}$  NMR Spectra Analysis (MiDSASA). *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1489–1496.
- (28) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK Users' Guide*, 3rd ed.; Society for Industrial and Applied Mathematics: Philadelphia, PA, 1999.
- (29) Anthony, M.; Bartlett, P. In *Neural Network Learning: Theoretical Foundations*; Cambridge University Press: Cambridge, U.K., 1999.
- (30) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. In *Numerical Recipes: The Art of Scientific Computing*; Cambridge University Press: Cambridge, U.K., 2007.

CI700256N