# Computer Assisted Simulation of $^{13}C$ Nuclear Magnetic Spectra of Monosaccharides

Brooke E. Mitchell and Peter C. Jurs*

Department of Chemistry, 152 Davey Laboratory, Penn State University, University Park, Pennsylvania 16802

Received July 6, 1995[⊗]

Mathematical models are developed that relate the structures of monosaccharides to their $^{13}C$ nuclear magnetic resonance spectra. The data set of monosaccharides consists of 55 monosaccharides in the six-membered ring configuration and 56 monosaccharides in the five-membered ring configuration. The structural environment of each carbon atom in the data set is encoded using numerical atom-based descriptors which are then used to develop linear regression models relating the $^{13}C$ chemical shift to the structural features. The atom-based descriptors used in this study encode topological, geometric, and electronic information about the carbon atoms in monosaccharides. Multiple linear regression analysis is used to develop an 11-descriptor model to predict the chemical shifts of pyranoses and pyranosides and an eight-descriptor model to predict the chemical shifts of furanoses and furanosides. The models are then submitted to computational neural networks, giving improved results with final training set rms errors of 1.03 ppm for pyranoses and pyranosides and 1.58 ppm for furanoses and furanosides.

## INTRODUCTION

Monosaccharides are the basic building blocks of biologically and industrially important molecules such as carbohydrates. The ability to determine the structure of these compounds is important, and one of the most common analytical tools used for structural identification of organic compounds is carbon-13 nuclear magnetic resonance ($^{13}C$ NMR) spectroscopy. It is useful, therefore, to be able to simulate the $^{13}C$ NMR spectra of such molecules.

One method of spectral simulation involves developing mathematical models that relate the $^{13}C$ chemical shift of an atom to its structural environment. This type of relationship is a quantitative structure−property relationship (QSPR) and is based on the assumption that the chemical shift of a particular atom depends on the environment in which the atom exists. The chemical shift provides a measure of how much a nucleus is shielded by electrons from an applied magnetic field. Individual carbon nuclei experience a magnetic field that is slightly different from the applied magnetic field based on the electronic environment surrounding it, and the electronic environment is determined by the structure of the molecule. Carbons in different chemical environments will therefore have different chemical shifts. The QSPR model is developed from a data set of compounds with experimentally determined chemical shifts, and it can then be used to predict chemical shifts for carbon atoms of compounds that were not used during model development. Recently QSPR models have been developed that relate the chemical shifts of quinolines and isoquinolines,[1] keto steroids,[2] and tetrahydropyrans,[3] among others, to their molecular structure.

The parametric approach employed involves relating the experimentally determined chemical shifts to atom-based descriptors. Descriptors are numerical representations of structural features of molecules, and atom-based descriptors attempt to encode the chemical environment of carbon atoms. The descriptors can be topological, geometric, or electronic in nature. Linear models are developed that relate carbon chemical shifts to atom-based descriptors. The models are of the form

$$S = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \qquad (1)$$

where $S$ is the chemical shift, $\beta_n$ is the linear regression coefficient, and $X_n$ is the descriptor value.

Improvement of the models is achieved with the use of computational neural networks, which can be considered as a nonlinear mathematical function. Improvements occur both because of the nonlinear nature and because of the larger number of adjustable parameters. Although the results of neural networks are better, the computational time required is much higher than for linear regression analysis. Regression coefficients for a given subset of descriptors can be found in a single step, while the optimization of the neural network adjustable parameters is an iterative process. The optimization of the neural network is performed using a quasi-Newton BFGS (Broyden−Fletcher−Goldfarb−Shanno)[4−8] algorithm, as opposed to the more commonly used back propagation algorithm,[9] because it has proven to give more accurate results and is more computationally efficient. The quasi-Newton BFGS algorithm utilizes second derivative information of the error function to determine the direction of a series of line minimizations. It does not have a learning rate or a momentum parameter that must be chosen by the user as in the back propagation algorithm.

## EXPERIMENTAL SECTION

The flow chart in Figure 1 shows the procedure used in this study. The software package used, Automated Data Analysis and Pattern Recognition Toolkit (ADAPT),[10,11] contains most of the programs used in the development of the QSPR models which were run on a Sun 4/110 workstation in our laboratory at Penn State University. The genetic algorithm program, the simulated annealing program,[12] and all computational neural networks[13] were run on a DEC 3000 AXP Model 500 workstation, also at Penn State University.

The data sets used in this study consisted of 55 monosaccharides in the six-membered ring configuration (33 pyra-

[⊗] Abstract published in *Advance ACS Abstracts,* December 1, 1995.

0095-2338/96/1636-0058$12.00/0 

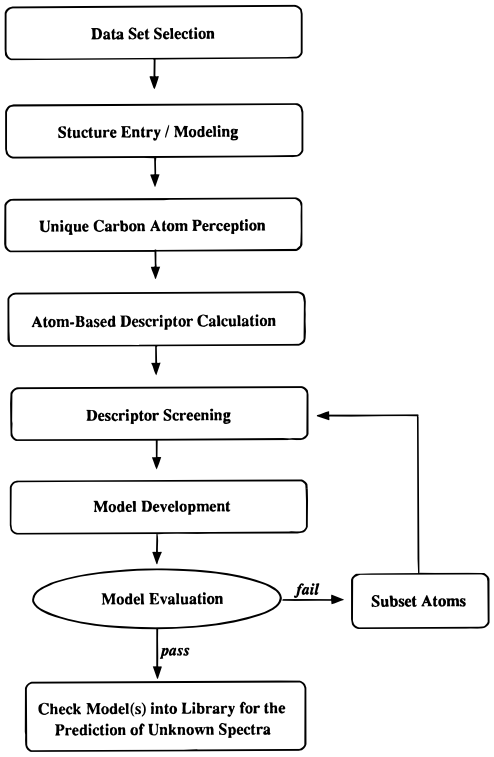


**Figure 1.** Flow chart of the ADAPT methodology used to develop spectral prediction models.

noses and 22 pyranosides) and 56 monosaccharides in the five-membered ring configuration (28 furanoses and 28 furanosides). These two groups will be referred to as the pyranoses and furanoses hereafter. Separate models were developed for the two monosaccharide subsets, as it was discovered that a single model could not adequately simulate the chemical shifts of a combined data set. The chemical shifts for the compounds were taken from 14 literature references[14−27] spanning a 17 year time range. The large number of references was necessary to provide as many spectra as possible in order to develop statistically strong models. Because NMR chemical shifts of monosaccharides are very dependent on the conditions under which they are taken (solvent, temperature, concentration, etc.) and the fact that NMR spectrometers change over a 17 year span, a calibration method described by McIntyre and Small[28] was used to standardize the literature data. The method involves choosing a set of standard experimental conditions and then calibrating the rest of the reference data to those conditions using linear regression. The main requirement for performing this calibration is that the $^{13}$C NMR spectra of some compounds exist under both the standard experimental conditions and the differing conditions of the literature data. The calibration between the chosen standard conditions and other conditions is performed by relating the chemical shifts of the compounds experimentally collected under both conditions via an equation of the form

$$S_{std} = C_0 + C_1 S_{lit} \quad (2)$$

where $S_{std}$ and $S_{lit}$ are the chemical shifts of carbon atoms under the chosen standard experimental conditions and reported literature conditions and $C_0$ and $C_1$ are simple regression coefficients. $C_0$ and $C_1$ which relate $S_{std}$ and $S_{lit}$ are determined for each different literature reference. Separate standard conditions were chosen for the pyranose data and the furanose data. The average $C_0$ value for pyranoses was 0.499 ppm with a range of −0.696 to 1.71 ppm and the average $C_1$ value was 0.996 with a range of 0.991 to 0.998. The average standard deviation of the linear models was 0.336 ppm. Similarly, the average $C_0$ value for furanoses was −0.412 ppm, ranging from −1.56 to 0.456 ppm and the average $C_1$ value was 0.997, ranging from 0.983 to 1.00. The average standard deviation of these models was 0.850 ppm.

The compounds in the data set were divided into training sets and prediction sets. The training set was used for model development, and approximately 10% of the compounds were held aside as an external prediction set which was used for model validation. The pyranose training set consisted of 49 compounds with the remaining six compounds in the prediction set. The furanose training set consisted of 50 compounds with an external prediction set of six compounds. The compounds used in this study and their references are listed in Table 1. For the computational neural networks used in this study, the training sets were further subdivided into a smaller training set and a cross-validation set. A cross-validation set was used to prevent overtraining of the network, as discussed below. The cross-validation set for the pyranoses consisted of five compounds, leaving a training set of 44 compounds, and the cross-validation set for the furanoses consisted of five compounds, leaving 45 compounds in the training set. The linear regression prediction sets were retained as computational neural network prediction sets for comparison purposes. The prediction sets and the cross-validation sets were chosen to be representative of the types of compounds in the entire data set, but the specific compounds were randomly chosen. The prediction sets and cross-validation sets are denoted by superscripts in Table 1.

Once the data set was compiled, the two-dimensional structures were sketched into the computer on a graphics terminal and stored as connection tables. Because accurate three-dimensional structures were necessary for geometric and electronic descriptor calculation, the molecular modeling routine, MM2,[29,30] was used to put the compounds in their energy minimized conformations. Although it is known that the $^{13}$C NMR spectrum of a monosaccharide results from the average motion of the compound in many different conformations, it was necessary for the compound to be in a fixed conformation for descriptor calculation so the lowest strain energy conformation was chosen, and since accurate results were obtained, further work in this area was not attempted.

In order to prevent models from being unduly influenced toward any one kind of carbon atom, only one carbon atom of each type was used during model development. The ADAPT software perceived and activated one carbon atom in each unique chemical environment for use in the study. The unique carbon atom perception resulted in a regression training set of 299 atoms for both the pyranoses and furanoses and in regression prediction sets of 37 atoms for the pyranoses and 34 atoms for the furanoses. The cross-validation sets taken from the training sets for computational neural networks consisted of 30 pyranose atoms and 29 furanose atoms.

ADAPT routines were used to calculate atom-based descriptors to accurately encode the chemical environment of the carbons in the monosaccharides. The descriptors were calculated directly from the energy minimized three-

**Table 1.** Monosaccharides Used and the References from Which Their $^{13}C$ NMR Spectra Were Obtained

| no. | compound name | ref | no. | compound name | ref |
|---|---|---|---|---|---|
| 1 | α-L-sorbopyranose | 19 | 57 | β-D-talofuranose | 23 |
| 2 | β-D-xylopyranose | 17 | 58 | α-D-ribofuranose | 24 |
| 3 | α-D-xylopyranose[b] | 17 | 59 | β-D-ribofuranose[a] | 24 |
| 4 | α-L-rhamnopyranose | 19 | 60 | β-D-fructofuranose | 21 |
| 5 | β-L-rhamnopyranose | 11 | 61 | α-D-fructofuranose[a] | 21 |
| 6 | α-D-glucopyranose | 17 | 62 | methyl-α-D-fructofuranoside[a] | 18 |
| 7 | β-D-glucopyranose[b] | 17 | 63 | α-D-fucofuranose | 19 |
| 8 | α-D-mannopyranose | 15 | 64 | β-D-fucofuranose | 19 |
| 9 | β-D-mannopyranose | 15 | 65 | α-D-arabinofuranose | 12 |
| 10 | α-D-galactopyranose[a] | 17 | 66 | β-D-arabinofuranose | 12 |
| 11 | β-D-galactopyranose[a] | 17 | 67 | methyl-β-D-ribofuranoside | 22 |
| 12 | α-D-talopyranose[a] | 17 | 68 | α-L-sorbofuranose | 21 |
| 13 | β-D-talopyranose | 17 | 69 | β-L-sorbofuranose | 21 |
| 14 | methyl-β-D-mannopyranoside | 15 | 70 | β-D-tagatofuranose | 21 |
| 15 | α-D-arabinopyranose | 17 | 71 | methyl-α-L-arabinofuranoside | 16 |
| 16 | β-D-arabinopyranose | 17 | 72 | methyl-β-D-fructofuranoside | 18 |
| 17 | α-D-lyxopyranose | 13 | 73 | methyl-α-L-sorbofuranoside | 18 |
| 18 | methyl-α-D-lyxopyranoside | 13 | 74 | methyl-β-L-sorbofuranoside | 18 |
| 19 | β-D-ribopyranose | 13 | 75 | methyl-α-D-tagatofuranoside | 18 |
| 20 | methyl-α-D-xylopyranoside | 13 | 76 | methyl-β-D-tagatofuranoside | 18 |
| 21 | methyl-β-D-xylopyranoside | 13 | 77 | β-D-galactofuranose | 19 |
| 22 | methyl-α-D-arabinopyranoside | 13 | 78 | 6-methyl-β-D-tagatofuranoside | 18 |
| 23 | methyl-β-D-arabinopyranoside[b] | 13 | 79 | α-D-psicofuranose | 18 |
| 24 | methyl-α-D-ribopyranoside[a] | 13 | 80 | β-D-psicofuranose | 18 |
| 25 | methyl-β-D-ribopyranoside | 13 | 81 | 6-methyl-α-D-psicofuranoside | 18 |
| 26 | methyl-α-D-glucopyranoside | 17 | 82 | 6-methyl-β-D-psicofuranoside[b] | 18 |
| 27 | methyl-β-D-glucopyranoside | 17 | 83 | α-D-idofuranose[b] | 20 |
| 28 | methyl-α-D-galactopyranoside | 17 | 84 | β-D-idofuranose | 20 |
| 29 | methyl-β-D-galactopyranoside | 17 | 85 | α-D-threofuranose | 24 |
| 30 | methyl-α-D-mannopyranoside[a] | 14 | 86 | β-D-erythrofuranose | 24 |
| 31 | β-D-allopyranose | 15 | 87 | α-D-altrofuranose | 24 |
| 32 | methyl-α-D-altropyranoside | 14 | 88 | β-D-altrofuranose | 24 |
| 33 | 2-methyl-α-D-glucopyranoside | 14 | 89 | β-D-threofuranose[b] | 24 |
| 34 | 2-methyl-β-D-glucopyranoside | 14 | 90 | methyl-α-D-erythrofuranoside[b] | 16 |
| 35 | 2-methyl-α-D-mannopyranoside[b] | 14 | 91 | methyl-β-D-erythrofuranoside | 16 |
| 36 | 2-methyl-β-D-mannopyranoside | 14 | 92 | methyl-α-L-threofuranoside | 16 |
| 37 | α-D-ribopyranose | 15 | 93 | methyl-β-L-threofuranoside | 16 |
| 38 | β-D-fructopyranose | 17 | 94 | methyl-β-L-arabinofuranoside[a] | 16 |
| 39 | methyl-α-idopyranoside | 15 | 95 | methyl-α-D-lyxofuranoside | 16 |
| 40 | β-D-lyxopyranose | 15 | 96 | methyl-β-D-lyxofuranoside | 16 |
| 41 | 3-methyl-α-D-glucopyranoside | 11 | 97 | methyl-α-D-ribofuranoside | 16 |
| 42 | 3-methyl-β-D-glucopyranoside | 11 | 98 | methyl-β-D-xylofuranoside | 16 |
| 43 | 1,3-dimethyl-β-D-galactopyranoside | 23 | 99 | methyl-α-D-allofuranoside | 16 |
| 44 | β-L-fucopyranose | 11 | 100 | methyl-α-D-galactofuranoside | 16 |
| 45 | α-D-tagatopyranose | 19 | 101 | methyl-β-D-galactofuranoside[b] | 16 |
| 46 | α-D-fucopyranose | 17 | 102 | methyl-α-D-glucofuranoside | 16 |
| 47 | β-D-fucopyranose[b] | 17 | 103 | methyl-β-D-glucofuranoside | 16 |
| 48 | α-D-gulopyranose | 24 | 104 | methyl-α-D-mannofuranoside | 16 |
| 49 | β-D-gulopyranose | 24 | 105 | methyl-β-D-mannofuranoside | 16 |
| 50 | α-D-altropyranose | 24 | 106 | α-D-erythrofuranose[a] | 24 |
| 51 | β-D-altropyranose | 24 | 107 | α-L-arabinofuranose | 19 |
| 52 | α-D-allopyrnose | 24 | 108 | β-L-arabinofuranose | 19 |
| 53 | α-L-fucopyranose[a] | 11 | 109 | β-D-allofuranose | 12 |
| 54 | α-L-arabinopyranose | 15 | 110 | α-D-talofuranose[a] | 24 |
| 55 | β-L-arabinopyranose | 15 | 111 | α-D-lyxofuranose | 12 |
| 56 | methyl-α-D-xylofuranoside | 16 | | | |

[a] Prediction set compound. [b] Cross-validation set compound.

dimensional structures. Topological descriptors calculated include simple and valence corrected connectivity indices and counts of atom types a certain number of bonds from the carbon center of interest. Geometric descriptors included powers of intermolecular distances between the carbon center of interest and oxygens as well as torsional angle descriptors. The electronic environment of each carbon center was described by atomic charge descriptors. The resulting descriptor pool was screened to ensure that only information rich descriptors were used for model development. Pairwise correlations were examined so that only one descriptor was retained from a pair contributing similar information (correlation coefficients $\geq 0.93$), and descriptors with greater than 80% identical values were dropped since those descriptors are not encoding the structural differences between monosaccharides that account for their chemical shift differences. This step of descriptor screening was an objective, as opposed to subjective, feature selection since decisions were made without knowledge or use of the dependent variable.

The goal of model development was to find a subset of the descriptors that accurately represented the chemical shifts. A variety of methods was used to generate linear regression models. Leaps-and-bounds regression analysis[31] calculates the best subsets of descriptors for varying subset size based on the $R^2$ criterion. A genetic algorithm routine and a

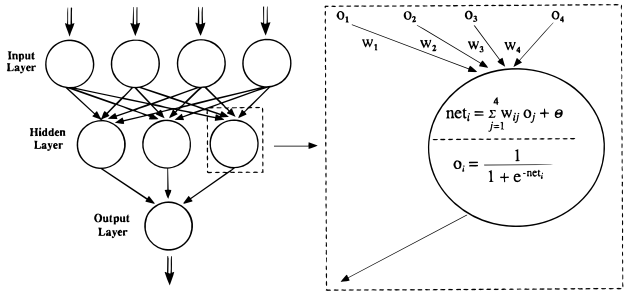


**Figure 2.** A typical computational neural network architecture.

simulated annealing routine, both based on RMS error, were also used to optimize the subset of descriptors. Both of these algorithms were developed in this research group. None of the three routines tested the statistical integrity of the models that were developed, so an interactive regression analysis routine was used for model validation. If the models found using these techniques are deemed to be unacceptable and better models cannot be found, the data set can be broken up into smaller, more homogeneous subsets, but this was not necessary beyond the division into five- and six-membered ring compounds. The best models found were an 11-descriptor model from the genetic algorithm routine for the pyranoses and an eight-descriptor model from the simulated annealing routine for the furanoses. Although the best pyranose model was found using the genetic algorithm routine and the best furanose model was found using the simulated annealing routine, both routines were used in combination with leaps-and-bounds regression analysis and the interactive regression routine for developing models for the two data sets.

Computational neural networks were then used to give improved results over the multiple linear regression analysis. Figure 2 shows the neural network architecture used in this research, which was a fully-connected, feed-forward, three layer neural network. The input layer consisted of as many neurons as there were descriptors in the linear regression model. The number of neurons in the hidden layer was optimized, and there was one neuron in the output layer, the estimated chemical shift. The number of hidden layer neurons was considered to be optimized when the training set error did not significantly decrease with the addition of another neuron. A linear transformation of the descriptor values restricted them to the interval [0,1] and these values were then used as the input for the network. Each neuron was connected to all of the neurons in the layer above it, and there was a weight associated with each connection. The input value, $net_i$, of the hidden layer was the sum of the product of each input layer value and its weight added to a bias term as shown in the enlarged neuron of Figure 2. This value was then converted to the output, $O_i$, using the nonlinear sigmoidal function shown in Figure 2

$$O_i = 1/(1 + e^{-net_i}) \qquad (3)$$

The outputs from the hidden layer then became inputs for the output layer and were treated similarly. The result from the output layer was a number on the interval [0,1] which was transformed to obtain the estimated carbon chemical shift. The network was trained to relate the descriptor values of the input layer to the associated chemical shift by iteratively adjusting the weights and biases to minimize the sum squared error between the observed and estimated chemical shifts. A cross-validation set was used to prevent overtraining of the network. The cross-validation set was a small subset randomly drawn from the training set which was not used during training but was tested periodically during training. When the cross-validation set error was minimized, training was stopped since beyond this point the network was fitting characteristics specific to training set compounds rather than general characteristics of the entire data set. The use of a cross-validation set increased the confidence with which external predictions could be made using the trained network.

The carbon chemical shifts were estimated and then assembled into the simulated $^{13}$C NMR spectra and compared with the observed spectra. The new models were added to the existing model library and used to predict the chemical shifts of the prediction set compounds.

## RESULTS AND DISCUSSION

In order to determine whether a need existed for a model to simulate the carbon chemical shifts of monosaccharides, 57 monosaccharide atoms were submitted to the ADAPT spectral prediction routine. This routine submits compounds to previously developed models and chooses the model that will best predict the carbon chemical shift. The standard error of estimate for the submitted atoms was 38.70 ppm, a clearly unacceptable error for spectra whose peaks cover a range of less than 200 ppm, so the need for a monosaccharide study was demonstrated.

The best model that could be developed for a data set consisting of a mixture of the five- and six-membered ring compounds had a training set rms error of 2.05 ppm, which was much higher than desired. Therefore the study was divided into two parts, developing separate models for the five-membered ring compounds and the six-membered ring compounds. Numerous models were developed for both sets of compounds using the previously mentioned linear regression routines, and the quality of these models was evaluated using the multiple correlation coefficient ($R$), the rms error, and the number of variables in the model.

The best models found for the pyranoses and the furanoses are listed in Table 2. The best 11-descriptor model for the pyranoses was found using the genetic algorithm routine. Models with a greater number of descriptors were either statistically unsound or did not significantly improve the chemical shift error or the $R$ value. The rms error of the training set for this linear regression model was 1.22 ppm with $R = 0.991$, and the rms error of the external prediction set compounds was 1.59 ppm with $R = 0.988$. Of the 11 descriptors in the model, three were topological (ACON 3, ICNC 4, and NNCA 1), one was electronic (WHK3 1), and the remaining seven were geometric (CXVD 1, COD3 1, HOD3 1, HXI3 2, HXT3 2, NHTR 60, and ODIS 1). The descriptor definitions are also listed in Table 2. Many of the geometric descriptors contained information about interactions between the carbon center of interest and oxygen atoms in the molecule. This was reasonable since oxygens are strong electron withdrawing atoms and their presence significantly affects the carbon chemical shift value, and there were multiple oxygens in the monosaccharide molecules. It was also not surprising that the majority of descriptors contained geometric information since the main differences

**Table 2.** Final Models for the Simulation of $^{13}C$ NMR Chemical Shifts of Pyranoses/Pyranosides and Furanoses/Furanosides

| coeff | sd of coeff | label | descriptor definition |
|---|---|---|---|
| | | | Pyranose/Pyranoside Model[b] |
| −22.0 | 2.55 | ACON 3 | av connectivity index three bonds from C* [a] |
| −3.40 | 0.438 | ICNC 4 | corrected connectivity index four bonds from C* |
| 3.54 | 0.282 | WHK3 1 | no. carbons one bond from C* |
| −12.7 | 1.40 | WHK3 1 | square root of the absolute value of the weighted sum of Hückel charges on heavy atoms three bonds from C* |
| −2.70 | 0.504 | CXVD 1 | van der Waals energy of C* interacting with all other heavy atoms |
| 104 | 1.70 | COD3 1 | sum of the inverse cubed throughspace distances from C* to all oxygens one bond from C* |
| −21.0 | 1.98 | HOD3 1 | square root of sum of the inverse cubed throughspace distances from hydrogens attached to C* to all oxygens one bond from C* |
| −136 | 6.19 | HXI3 2 | sum of the inverse cubed throughspace distances from hydrogens attached to C* to all heavy atoms four bonds from C* |
| 89.9 | 2.94 | HXT3 2 | sum of the inverse cubed throughspace distances from hydrogens attached to C* to all heavy atoms from three to four bonds from C* |
| 0.590 | 0.098 | NHTR 60 | no. of 60° hydrogen torsional angles involving C* |
| 55.8 | 3.43 | ODIS 1 | inverse throughspace distance from C* to the nearest oxygen |
| 6.82 | 2.95 | | Y intercept |
| | | | Furanose/Furanoside Model |
| 140 | 6.66 | ACNC 2 | av corrected connectivity index two bonds from C* |
| −2.86 | 0.206 | AVC2 2 | no. secondary heavy atoms two bonds from C* |
| 8.49 | 0.938 | WPAT 1 | sum of weighted paths originating from C* |
| 103 | 2.90 | MNAC 1 | most negative atomic charge on heavy atoms one bond from C* |
| 93.1 | 1.79 | TOAC 1 | sum of the absolute values of atomic charges for all heavy atoms one bond from C* |
| 38.5 | 6.44 | HOD3 2 | sum of the inverse cubed throughspace distances from hydrogens attached to C* to all oxygens two bonds from C* |
| 4.32 | 0.753 | HOEL 1 | van der Waals energy of hydrogens attached to C* interacting with all oxygen atoms three or more bonds from C* |
| −467 | 12.7 | SMBS 1 | length of shortest bond attached to C* |
| 663 | 18.6 | | Y intercept |

[a] C* is the carbon center for which the descriptor is being calculated. [b] $R = 0.996$, rms error = 1.22 ppm, $n = 299$ atoms. [c] $R = 0.992$, rms error = 1.91 ppm, $n = 299$ atoms.

between the pyranose compounds was in the geometric positions of the atoms. The compounds were very similar both topologically and electronically.

The same 11 descriptors were then submitted to computational neural networks with 11 input neurons and one output neuron. The number of hidden layer neurons was found to be optimum at five neurons. A greater number of hidden layer neurons did not give an improved training set rms error. An automated version of the quasi-Newton BFGS training algorithm was used to perform 500 neural network training sessions using different random starting weights, since the results are dependent upon the starting weights. A cross-validation set was used to determine when to stop training in order to prevent overtraining. The computational neural networks resulted in improved training set, cross-validation set, and prediction set rms errors of 1.03, 0.766, and 1.11 ppm, respectively. The improvement in results over linear regression was probably due both to the fact that the neural network used a larger number of adjustable parameters and the fact that it utilized a nonlinear mathematical function. Calculated and predicted chemical shift vs observed chemical shift plots for the pyranose neural network results are shown in Figures 3 and 4.

The simulated annealing algorithm found the best eight-descriptor model for the furanoses. Once again, models with a greater number of descriptors were either statistically unsound or did not improve the error or the $R$ value. The model contains three topological descriptors (ACNC 2, AVC2 2, and WPAT 1), two electronic descriptors (MNAC 1 and TOAC 1), and three geometric descriptors (HOD3 2, HOEL 1, and SMBS 1). The geometric descriptors in this model also involve the interaction between the carbon center and nearby oxygen atoms. The training set rms error was 1.91 ppm with $R = 0.992$ and the prediction set rms error was 1.23 ppm with $R = 0.997$. The fact that the prediction set had a much lower error than the training set indicated that the model accurately encoded the structural features that

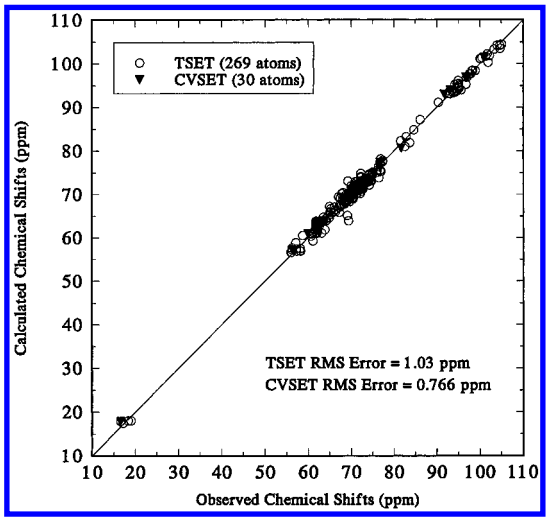


**Figure 3.** Calculated versus observed chemical shift values for the training (TSET) and cross-validation (CVSET) sets from the 11 descriptor pyranose model after submission to computational neural networks.

determine chemical shift for the prediction set compounds.

The eight descriptors found by linear regression for the furanoses were then submitted to computational neural networks. The optimal network architecture was found to be eight input neurons, five hidden layer neurons, and one output neuron. The best results found using the automated version of the quasi-Newton BFGS training algorithm gave training set, cross-validation set, and prediction set rms errors of 1.58, 0.995, and 0.898 ppm, respectively. Improvement was seen over regression results, with the prediction set error still being lower than the training set error. Figures 5 and 6 show the calculated and predicted chemical shift vs observed chemical shift plots for the furanose neural network results.

The accuracy of the models that have been developed can be seen in Figure 7, which compares the simulated and observed spectra of α-L-arabinopyranose as an example. The

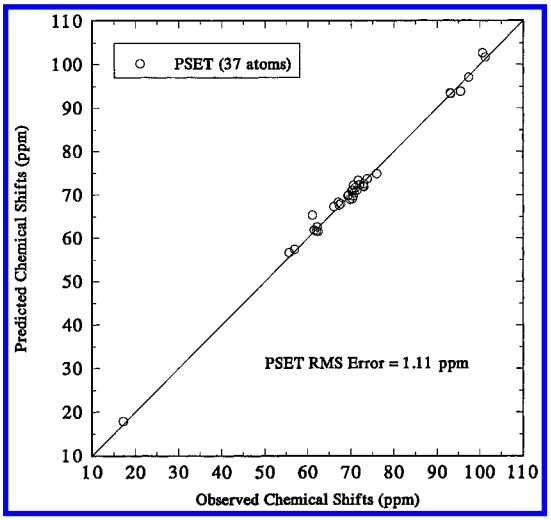


**Figure 4.** Validation of the 11 descriptor pyranose model after submission to computational neural networks using the external prediction set (PSET).

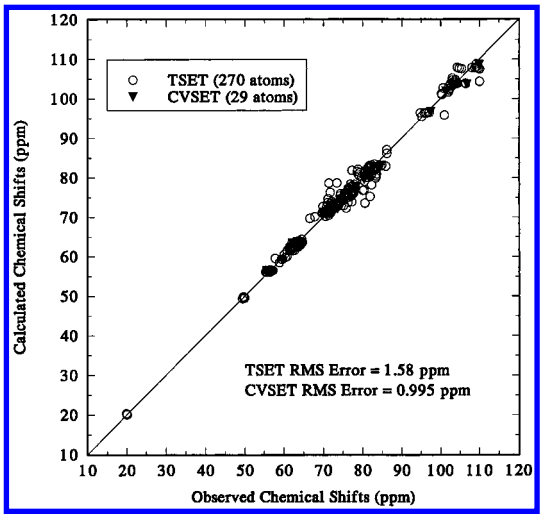


**Figure 5.** Calculated versus observed chemical shift values for the training (TSET) and cross-validation (CVSET) sets from the eight descriptor furanose model after submission to computational neural networks.

chemical shifts in these spectra were obtained with the linear regression pyranose model. The overall spectral error between the two spectra was 0.480 ppm, and they were virtually indistinguishable by visual inspection.

Both data sets contained a small number of chemical shifts that were significantly different from the main group of chemical shifts, and the possibility that these points were highly leveraged in the model was considered. The shifts were removed, and the statistics of the model were recalculated. In both models it was found that the coefficients of the descriptors and the model statistics did not significantly change with the removal of the questionable shifts so the models were deemed unleveraged by these shifts.

## CONCLUSION

Multiple linear regression models have been developed with the use of the ADAPT methodology for the simulation of $^{13}$C NMR spectra of monosaccharides. The model for the five-membered ring configuration compounds is the first that has been developed, and the model for the six-membered ring configuration compounds gives comparable results to previous work by Small and McIntyre[28] using similar

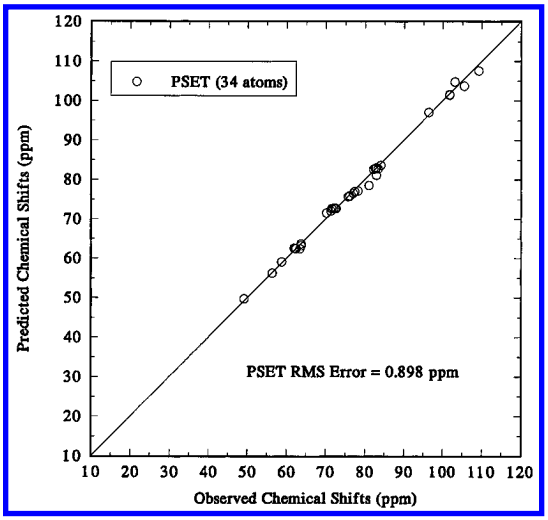


**Figure 6.** Validation of the eight descriptor furanose model after submission to computational neural networks using the external prediction set (PSET).

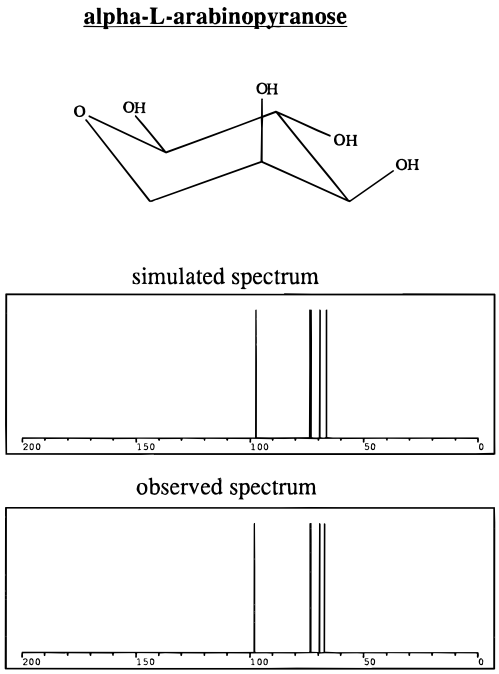


**Figure 7.** Simulated and observed spectra for α-L-arabinopyranose.

methodology without needing to subset the carbon atoms into smaller, more homogeneous groups. Small and McIntyre developed five separate models to simulate the NMR chemical shifts of different types of carbon atoms in pyranoses. The results of these models are improved upon with the use of computational neural networks. The pyranose and furanose training set errors improved by 16% and 17%, respectively, and the prediction set errors improved by 30% and 27%, respectively, over the linear regression results. The models have been shown to accurately predict chemical shifts of compounds not used during their development.

## REFERENCES AND NOTES

(1) Ranc, M. L.; Jurs, P. C. Simulation of carbon-13 nuclear magnetic resonance spectra of quinolines and isoquinolines. *Anal. Chim. Acta* **1991**, *248*, 183−193.

(2) Sutton, G. P.; Anker, L. S.; Jurs, P. C. Evaluation of automated methods for the selection of models for the simulation of carbon-13 nuclear magnetic resonance spectra of keto steroids. *Anal. Chem.* **1991**, *63*, 443−449.

(3) Clouser, D. L.; Jurs, P. C. Simulation of $^{13}C$ nuclear magnetic resonance spectra of tetrahydropyrans using regression analysis and neural networks. *Anal. Chim. Acta* **1994**, *295*, 221−231.
(4) Broyden, C. G. The convergence of a class of double-rank minimization algorithms. *J. Inst. Maths. Appl.* **1970**, *6*, 76.
(5) Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, 13, 317.
(6) Goldfarb, D. A. family of variable-metric methods derived by variational means. *Math. Comput.* **1970**, *24*, 23.
(7) Shanno, D. F. Conditioning of quasi-Newton methods for function minimizations. *Math. Comput.* **1970**, *24*, 647.
(8) Fletcher, R. *Practical Methods of Optimization, Vol. 1, Unconstrained Optimization*; Wiley: New York, 1980.
(9) Jansson, P. A. Neural networks: An overview. *Anal. Chem.* **1991**, *63*, 357A−362A.
(10) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(11) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103−129.
(12) Sutter, J. M.; Jurs, P. C. Selection of molecular structure descriptors for quantitative structure-activity relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B.V.: Amsterdam, 1995.
(13) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative structure-activity relationships for toxicity of phenols using regression analysis and computational neural networks. *Environmental Toxicol. Chem.* **1994**, *13(5)*, 841−851.
(14) Dorman, D. E.; Roberts, J. D. Nuclear magnetic resonance spectroscopy. Carbon spectra of some pentose and hexose aldopyranoses. *J. Am. Chem. Soc.* **1970**, *92(5)*, 1355−1361.
(15) Bock, K.; Pedersen, C. Carbon-13 nuclear magnetic resonance spectroscopy of monosaccharides. *Adv. Carbohydr. Chem. Biochem.* **1983**, *41*, 27−66.
(16) Bock, K.; Pedersen, C. A study of $^{13}CH$ coupling constants in pentopyranoses and some of their derivatives. *Acta. Chem. Scand. B* **1975**, *29*, 258−264.
(17) Bock, K.; Pedersen, C. A study of $^{13}CH$ coupling constants in hexopyranoses. *J. Chem. Soc., Perkin Trans. 2* **1974**, *3*, 293−297.
(18) Perlin, A. S.; Casu, B.; Koch, H. J. Configurational and conformational influences on the carbon-13 chemical shifts of some carbohydrates. *Can. J. Chem.* **1970**, *48*, 2596−2606.
(19) Ritchie, R. G. S.; Cyr, N.; Korsch, B.; Koch, H. J.; Perlin, A. S. Carbon-13 chemical shifts of furanosides and cyclopentanols. Configurational and conformational influences. *Can. J. Chem.* **1975**, *53*, 1424−1433.
(20) Pfeffer, P. E.; Valentine, K. M.; Parrish, F. W. Deuterium-induced differential isotope shift $^{13}C$ NMR. 1. Resonance reassignments of mono- and disaccharides. *J. Am. Chem. Soc.* **1979**, *101(5)*, 1265−1274.
(21) Angyal, S. J.; Bethell, G. S. Conformational analysis in carbohydrate chemistry. III* The $^{13}C$ NMR spectra of the hexuloses. *Aust. J. Chem.* **1976**, *29*, 1249−1265.
(22) Reuben, J. Fingerprints of molecular structure and hydrogen bonding effects in the carbon-13 nmr spectra of monosaccharides with partially deuterated hydroxyls. *J. Am. Chem. Soc.* **1984**, *106(21)*, 6180−6186.
(23) Reuben, J. Effects of solvent and hydroxyl deuteration on the carbon-13 nmr spectrum of D-idose: spectral assignments, tautomeric compositions, and conformational equilibria. *J. Am. Chem. Soc.* **1985**, *107(21)*, 5867−5870.
(24) Wolff, V. G.; Breitmaier, E. Carbon-13 NMR spectroscopic determination of the keto form of D-fructose, L-sorbose and D-tagatose in aqueous solutions. *Chem.-Ztg.* **1979**, *103*, 232−233.
(25) Breitmaier, E.; Jung, G.; Voelter, W. Fourier transform carbon-13 NMR spectroscopy. 12 Pulse Fourier transform carbon-13 NMR of ribofuranosides and ribopyranosides. *Chimia* **1972**, *26*, 136−139.
(26) Voelter, W.; Breitmaier, E. Influence of anomeric attached adenine residues on the carbon-13 chemical shifts of pyranoses. *Org. Magn. Reson.* **1973**, *5(7)*, 311−319.
(27) King-Morris, M. J.; Serianni, A. S. $^{13}C$ NMR studies of [1-$^{13}C$]aldoses: empirical rules correlating pyranose ring configuration and conformation with $^{13}C$ chemical shifts and $^{13}C$-$^{13}C$ spin couplings. *J. Am. Chem. Soc.* **1987**, *109(12)*, 3501−3508.
(28) McIntyre, M. K.; Small, G. W. Carbon-13 nuclear magnetic resonance spectrum simulation methodology for the structure elucidation of carbohydrates. *Anal. Chem.* **1987**, *59(14)*, 1805−1811.
(29) Burket, U.; Allinger, N. L. *Molecular Mechanics, ACS Monograph 177*; American Chemical Society: Washington, DC, 1982.
(30) Clark, T. A. *A Handbook of Computational Chemistry: A Practical Guide to Chemical Structure and Energy Calculations*; Wiley-Interscience: New York, 1985.
(31) Furnival, G.; Wilson, R. Regression by leaps and bounds. *Technometrics* **1974**, *16*, 499−511.

CI950262Y