

Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation

Peter C. Jurs,* Jon W. Ball, Lawrence S. Anker, and Todd L. Friedman

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

Received February 10, 1992

A combined approach using database retrieval and empirical modeling methods is described for simulation of ^{13}C nuclear magnetic resonance spectra directly from molecular structure. In this approach, predictive equations are retrieved from a library of regression models to simulate the ^{13}C NMR spectra of organic compounds. The enhancement of the regression model library is described by adding the capability to simulate accurate ^{13}C NMR shifts for side-chain atoms on cyclic compounds and for selected atoms in hydroaromatic compounds.

INTRODUCTION

Carbon-13 nuclear magnetic resonance spectrum simulation involves converting structural information into a simulated spectrum. Spectral simulation can be very useful in aiding the chemist in the solution of complex structure elucidation problems and in the verification of chemical shift assignments. The three most common methods of ^{13}C NMR spectral simulation involve linear additivity relationships, database retrieval techniques, and empirical modeling.

The utility of linear additivity relationships has been studied by many researchers including Zupan¹ and Fürst.² This technique utilizes a database of carbon atoms and their associated chemical shifts to derive empirical parameters for a variety of structures and functionalities. The chemical shift of a new atom in a particular environment can then be calculated using the empirical parameters derived from similar atomic environments in the form of simple linear equations. Although this approach can be applied to a wide range of compounds, standard errors in excess of 5 ppm are common.

Similarly, database retrieval techniques^{3,4} rely on access to a large database containing various carbon atoms and their corresponding known chemical shifts. In order to simulate the ^{13}C NMR spectrum for an unknown, the most similar atomic match is found from the database for each unique atom in the unknown. The chemical shift associated with each matched database atom is retrieved as the predicted chemical shift for each atom in the unknown. Although this method is applicable to a wide variety of compounds, the accuracy is highly dependent on the size and breadth of the database and in the similarity metric used.

Empirical modeling techniques involve the development of linear mathematical models from a data set of known chemical shifts. These models relate the chemical shift of an atom to various atom-based structural parameters (descriptors) and have the form

$$S = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n \quad (1)$$

where the S is the chemical shift of an individual carbon atom, the X_i 's are the descriptor values, and the b_i 's are coefficients determined by multiple linear regression analysis. Once developed, these models can be utilized to predict the chemical shifts for carbons not contained in the original data set. A limitation of this technique arises because the model equations can only be used to predict the shifts for atoms that are structurally similar to the atoms used to develop the models. Unlike database retrieval techniques, however, these models

have the ability to interpolate chemical shifts. In addition, this technique typically yields simulated spectra with standard errors on the order of 1.0 ppm.

The goal of this research effort is to develop the capability to simulate accurate ^{13}C NMR spectra for a wide variety of organic compounds directly from their structures in an automated fashion. In the past, our primary focus was in the development of empirical models for limited sets of structurally homogeneous compounds. However, more recently, we have been studying the utility of combining database retrieval and empirical modeling techniques in an attempt to benefit from the advantages offered by both techniques. In this combined approach, predictive equations are retrieved from a library of stored regression models to calculate the chemical shifts for each atom in a query compound. For this approach to be successful, two tasks must be accomplished. First, the library of predictive regression equations must be broadly applicable, that is, models capable of simulating accurate shifts for many different atomic environments must be developed and stored. Second, an effective algorithm capable of selecting the most appropriate model to use for predicting the shift of a query atom needs to be developed.

CONSTRUCTION OF THE REGRESSION MODEL LIBRARY

A total of 71 regression models capable of simulating ^{13}C NMR spectra for a variety of structural classes have been developed. They have been stored in a library of regression equations. The structural classes represented are as follows: linear and branched alkanes,⁵ cycloalkanes,⁶ cyclohexanols and decalols,⁷ hydroxysteroids,⁸ cyclopentanes and cyclopentanols,⁹ norbornanols,¹⁰ cyclohexanones and decalones,¹¹ piperidines,¹² polychlorinated biphenyls,¹³ alkyl-substituted benzenes and polyaromatics,¹⁴ ketosteroids,¹⁵ quinolines and isoquinolines,¹⁶ cyclopentanones and cycloheptanones.¹⁷ These models are listed in Table I.

As shown in Table I, three to eight models have been developed for each compound class. Each model was developed for a subset of the carbon atoms found in the compounds being studied, and the number of atoms used in the development of each model is shown in parentheses after the identity of each model. Models 8, 12, and 63 were each developed with less than eight observations. Because fewer than eight observations were not sufficient to generate statistically valid regression models, the average chemical shift of each atom subset was used in place of traditional regression equations.

Table I. Summary of Regression Model Library

description (no. of observations)	model no.	description (no. of observations)	model no.
	Linear/Branched Alkanes		
primary carbons (128)	1	tertiary carbons (53)	3
secondary carbons (119)	2	quaternary carbons (24)	4
	Cycloalkanes		
primary carbons (50)	5	tertiary carbons (79)	7
secondary carbons (157)	6	quaternary carbons (7)	8
	Cyclohexanols/Decalols		
primary carbons (39)	9	tertiary carbons (78)	11
secondary carbons (138)	10	quaternary carbons (8)	12
	Hydroxysteroids		
primary carbons (48)	13	tertiary with -OH (25)	16
secondary carbons (224)	14	quaternary carbons (53)	17
tertiary carbons (120)	15		
	Cyclopentanes/Cyclopentanols		
primary carbons (35)	18	all carbons with attached -OH (20)	23
secondary carbons (82)	19	tertiary and quaternary ring carbons (28)	24
tertiary carbons (36)	20	tertiary and quaternary side chain carbons (13)	25
tertiary with -OH (11)	21	secondary carbons (21)	26
quaternary carbons (13)	22	tertiary carbons with attached -OH (32)	27
	Norbornan-2-ols		
primary carbons (50)	28	tertiary carbons with attached -OH (30)	31
secondary carbons (95)	29	quaternary carbons (17)	32
tertiary carbons without attached -OH (82)	30	quaternary carbons without attached -OH (15)	33
	Cyclohexanones/Decalones		
primary carbons (46)	34	carbonyl carbons (38) (model B)	39
secondary carbons (170)	35	carbons one bond from C=O (67)	40
tertiary carbons (49)	36	carbons two bonds from C=O (85)	41
quaternary carbons (22)	37	carbons three or more bonds from C=O (135)	42
carbonyl carbons (38) (model A)	38		
	Piperidines		
all primary carbons (60)	43	primary carbons attached to N (16)	47
secondary carbons (78)	44	carbons one bond from N (62)	48
tertiary carbons (44)	45	carbons two bonds from N (65)	49
primary carbons attached to C (44)	46	carbons three or more bonds from N (55)	50
	Polychlorinated Biphenyls		
carbons with attached chlorines (84)	51	carbons with attached hydrogens (131)	53
carbons with attached C's, bridging carbons (49)	52		
	Alkyl-Substituted Benzenes/PAHs		
all aromatic ring carbons (231)	54	non-ring-bridging aromatic ring carbons (178)	56
methyl carbons (36)	55	ring-bridging aromatic ring carbons (53)	57
	Ketosteroids		
all carbonyl carbons (26)	58	all carbons 3 bond removed from the carbonyl (97)	61
all carbons 1 bond removed from the carbonyl (52)	59	all carbons 3bond removed from the carbonyl (208)	62
all carbons 2 bond removed from the carbonyl (74)	60		
	Quinolines/Isoquinolines		
all methyl carbons (5)	63	carbons 1 bond from bridge carbon OR 2 bonds from bridge carbon and nitro group (114)	65
carbons 2 bonds from bridge carbons and 2 bonds from nitro group (105)	64	bridge carbons and carbons attached to nitro group (69)	66
	Cyclopentanones/Cycloheptanones		
all carbonyl carbons (36)	67	secondary carbons 2 or more bonds from carbonyl (64)	70
carbons 1 bond removed from the carbonyl (64)	68	tertiary/quaternary carbons 2 or more bonds from carbonyl (43)	71
primary carbons 2 or more bonds from carbonyl (46)	69		

Each compound class was studied independently, and a standard set of procedures was followed in order to develop the regression models. The methodology used to develop these linear mathematical models has been described previously¹⁸ and, therefore, will only be briefly reviewed here.

Selecting a Data Set. The first step involved in a ¹³C NMR spectral simulation study is determining the class of compounds to be studied. Once a set of compounds is selected, their corresponding ¹³C NMR spectra are obtained from the primary literature. Primary literature data are used in order to minimize the probability of errors in the data. Typically 90% of the data set is used as the training set (a set used to generate the model equations), and the remaining 10% is set aside as a prediction set (a set used to test the external predictive ability of the model equations).

Structure Entry and Modeling. The structures are then entered as two-dimensional sketches and stored as connection tables in computer disk files. These two-dimensional representations are then submitted to a classical molecular mechanics routine in order to obtain the approximate energy-minimized three-dimensional coordinates. Three-dimensional representations of the structures are necessary for later computation of geometrical descriptors. In addition, the ¹³C NMR spectrum for each structure is stored.

Unique Atom Perception. It is important to include only the unique atoms (those in structurally unique surroundings) in the data set when generating models. Unique atoms give rise to unique chemical shifts. If an atom and its corresponding chemical shift are represented more than once in the data set, then the models can become unduly skewed or biased to

account for that particular atom type.

Atom Subsetting. The atoms in a given data set are usually structurally diverse, so it is difficult to generate one equation capable of simulating accurate ^{13}C NMR shifts for all the atoms together. Therefore, each data set must be divided into smaller, more homogeneous sets of atoms. In past studies, subsetting has been primarily based on connectivity (1° , 2° , 3° , 4°) or position relative to a functional group (e.g., the number of bonds from a carbonyl group). Model equations are then developed independently for each atom subset.

Descriptor Calculation. The atomic environment of each carbon atom is then calculated by numerically encoding topological, electronic, and geometrical features surrounding each atom. Topological descriptors include atom and valency counts as well as connectivity indexes. Electronic descriptors encode features based on partial atomic charges. These charges are computed using a variety of methods including extended Hückel and Del Re σ charge calculations. The geometrical descriptors encode information such as throughspace distances to other atoms in the molecule. Currently, the software supports the calculation of more than 750 descriptors.

Descriptor Screening. In order to assure their statistical significance, the descriptors are screened before being submitted to multiple linear regression analysis. Only information-rich descriptors pass the screening step onto regression analysis. This step is performed independently for each atom subset since a particular descriptor may contain useful information for one subset but not for another. Descriptors are generally removed from consideration if they contain mostly identical or zero values for the entire atom subset or if they are highly correlated with other descriptors.

Model Development. The descriptors that survive screening for each subset are submitted to multiple linear regression analysis. Forward selection and backward deletion are among several methods used to select descriptors. The objective is to find the best mathematical model that relates the chemical shift of a carbon atom to the available set of descriptors. A model is developed for each atom subset. After the models are developed, the simulated spectra can be assembled from the individual chemical shift predictions.

Model Evaluation. Various methods are utilized to validate each model. Models with high multiple correlation coefficients, R , and low standard errors of regression, s , are desirable. The goal is to find models with s values less than 1.0 ppm. Plots of calculated versus observed chemical shifts and residual error versus calculated shifts can be used to visually test for outliers, nonconstant variance, and other statistical problems. Each model is also internally validated using jackknifing. Another technique used to evaluate the models involves using the simulated spectra as probes into a library of stored spectra. If a simulated spectrum is of high quality, then the most similar spectrum (the spectrum retrieved as the top match) should be the observed spectrum of the same compound or a very similar one.

Another important model evaluation step is external prediction. Using an external prediction set, the models are used to predict the shifts for atoms which were not included in the original training set. Once a model has been validated, it is checked into the library so it can be retrieved for future spectral simulations.

MODEL SELECTION

To use the models for predicting the shifts of unknowns, an effective procedure for choosing the best model from the library

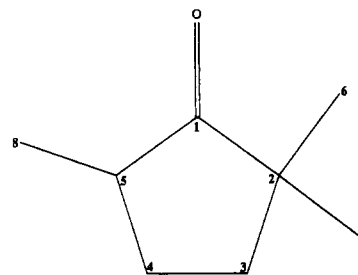


Figure 1. 2,2,5-Trimethylcyclopentanone with unique carbons labeled.

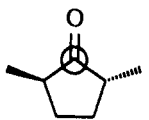
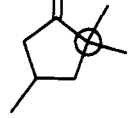
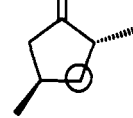
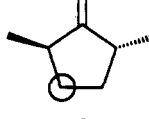
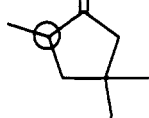
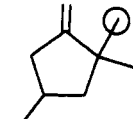
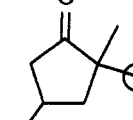
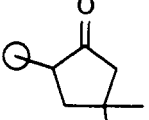
must be available. Currently, a nearest neighbor approach to model selection is being utilized. The development of this approach was recently described.¹⁷ The current model library contains 71 predictive equations which were developed using a total of 3896 atoms. The atomic environment surrounding each of these atoms is encoded using seven topological descriptors. These seven descriptors include the six EVEC descriptors¹⁹ used in the original model selection procedure²⁰ and Randić's atomic ID descriptor.²¹ To select a model for the prediction of a query atom in an unknown, the following procedure is followed. The seven descriptors are calculated for the query atom. In order to find the most similar atomic match, the Euclidean distance, d , between the query atom and each of the 3896 atoms in the database is calculated using eq 2, where the D_i 's are the descriptor values for an individual

$$d = [(D_1 - Q_1)^2 + (D_2 - Q_2)^2 + \dots + (D_7 - Q_7)^2]^{1/2} \quad (2)$$

database atom, and the Q_i 's are the descriptor values for the query atom. The database atom associated with the smallest Euclidean distance is defined as the nearest neighbor atom. In a database retrieval scheme, the shift associated with this nearest neighbor atom would be assigned to the query atom. In our scheme, the model developed from the nearest neighbor atom is used to calculate the chemical shift for the query atom. To do this, the descriptors that were used in the selected model are computed for the query atom. Then, these descriptor values are multiplied by the appropriate coefficients, and the chemical shift is predicted. This entire procedure is repeated for each atom in the query compound until the complete spectrum is simulated.

To illustrate how well this methodology works for compounds that are similar to the ones used to generate the 71 stored model equations, the spectrum for 2,2,5-trimethylcyclopentanone was simulated. This compound was not used to develop any models in the database, but it is similar to the compounds used in the cyclopentanone and cycloheptanone study. The structure for 2,2,5-trimethylcyclopentanone with its unique atoms numbered is shown in Figure 1. Table II shows the database atoms selected as the nearest neighbors, the Euclidean distance to each of them, and the shift prediction results using the models developed from the nearest neighbors. For example, Table II indicates that atom 1 in the query compound found a carbonyl carbon in a cyclopentanone molecule as the closest match in the database. Therefore, model 67, a carbonyl carbon model, was used to predict the shift for this atom. This was an appropriate selection. In each of the seven cases, an appropriate model was selected and very accurate predictions were obtained. Figure 2 illustrates visually the high degree of similarity between the simulated and observed spectra. The standard error of estimate for this simulated spectrum was 0.81 ppm. It should be emphasized that this entire procedure is fully automated, and it took less than 4 min to generate a complete simulated

Table II. Automated Prediction Results for 2,2,5-Trimethylcyclopentanone

atom	nearest neighbor	distance	model	obv. shift	pred. shift
1		0.1151	67	224.10	224.65
2		0.0152	68	44.70	44.24
3		0.0467	70	36.60	38.07
4		0.0789	70	28.00	27.93
5		0.0234	68	43.10	42.54
6		0.0118	69	24.10	24.80
7		0.0118	69	24.80	24.96
8		0.0184	69	15.20	15.05

spectrum starting from the energy-minimized three-dimensional structure.

This automated model selection procedure can be used to probe the regression model library systematically to determine if any weaknesses are present. A weakness in the library means that the ¹³C NMR spectra for a class of compounds cannot be simulated with the desired accuracy. When such deficiencies are identified, models which address them can be developed. Two independent studies were recently performed which probed the model library and identified two such weaknesses. The first study involved a set of cyclic compounds containing various side chains, and the second involved a set of hydroaromatic compounds.

CYCLIC COMPOUNDS CONTAINING SIDE CHAINS

Compounds I–IV are shown in Figure 3 with their unique side-chain atoms labeled. They were used to determine if a

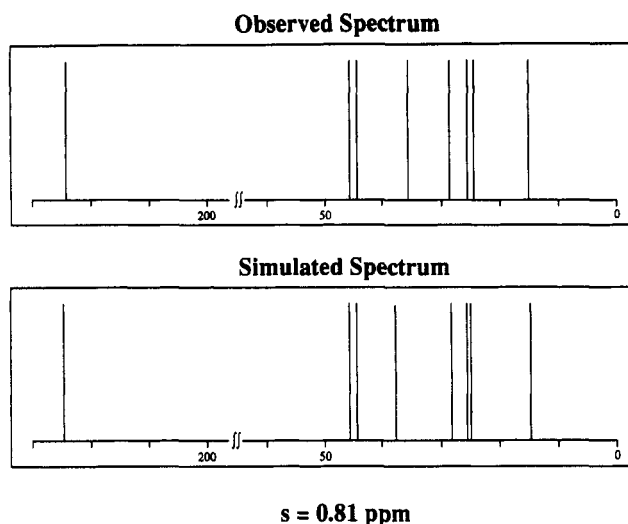


Figure 2. Simulated versus observed spectra for 2,2,5-trimethylcyclopentanone using models selected by the nearest neighbor approach.

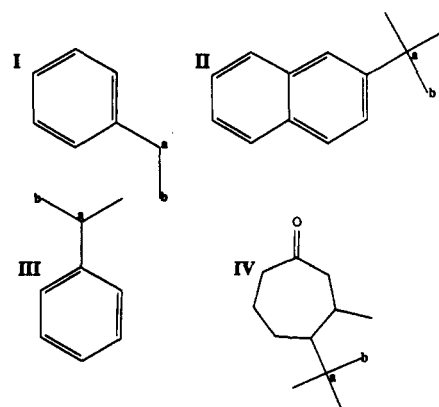


Figure 3. Probe compounds used to test if weaknesses existed in the regression model library.

weakness was present in the regression model library. Using the nearest neighbor model selection algorithm, models were selected from the database to predict the spectra for these four compounds. The prediction results using the selected models are summarized in Table III, Section A. The mean residual error for the side-chain atoms, atoms a and b in compounds I–IV, was 17.78 ppm, an unsatisfactory result. This is because models had not been developed for side-chain atoms on cyclic compounds, and therefore, the chemical shift predictions for these atoms were quite poor. It was reported earlier that the model database lacked the ability to simulate the shifts for atoms in side chains on cyclic compounds.¹⁷ The nearest neighbor distances, shown in Table III, Section A, are relatively large, indicating that the query atoms and the selected nearest neighbor atoms were in substantially different surroundings. The mean residual error for all the other atoms in each compound, however, was less than 1.00 ppm, which is satisfactory. The ring-carbon predictions were very accurate because models capable of simulating the shifts for benzenes, PAHs, and cycloheptanones have already been developed and are present in the model library.

Library searching was performed to further study the influence of the poor side-chain predictions on the overall spectra. The simulated spectra were compared to a library of about 1000 stored reference spectra. The observed spectra for compounds I–IV were among the library of reference spectra. Using a Euclidean distance metric, the top five spectral matches were retrieved from the spectral library for each simulated spectrum. The results of the library search

Table III. Prediction Results for Compounds I–IV

compd	atom a		atom b		mean residual other atoms, ppm	overall <i>s</i>
	nearest neighbor distance	residual error, ppm	nearest neighbor distance	residual error, ppm		
Section A. Before Side-Chain Models Were Present in Regression Model Library						
I	0.920	7.65	1.174	46.25	0.70	20.98
II	1.292	8.68	1.822	14.20	0.49	5.06
III	1.175	9.81	1.201	41.85	0.90	19.25
IV	0.846	4.68	0.780	9.13	0.78	3.52
Section B. After Side-Chain Models Were Present in Regression Model Library						
I	0.200	0.13	0.226	0.20	0.70	0.86
II	0.003	0.18	0.004	0.61	0.49	0.64
III	0.411	0.42	0.153	0.72	0.90	1.11
IV	0.119	0.38	0.130	1.40	0.78	1.09

Table IV. Library Searching Results for Simulated Spectra of Compounds I–IV

compd	rank	score
Section A. Not Using Side-Chain Models		
I	4	1090
II	2	281
III	3	1400
IV	1	65.4
Section B. Using Side-Chain Models		
I	1	2.8
II	1	4.1
III	1	3.9
IV	1	9.0

are listed in Table IV, Section A. The observed spectrum for only one compound, compound IV, was retrieved as the top spectral match. The score associated with this top match, 65.4, was quite high. The observed spectra for compounds I–III were also retrieved in the top five for each simulated spectrum. However, the scores associated with these matches were extremely high. Excessively high scores indicate that no spectrum in the database was very similar to the query spectrum, reducing the confidence in the retrievals. The score is a measure of the similarity between the simulated and the reference spectra and is computed by summing the squares of the errors between the shifts in the simulated spectrum and the shifts in the reference spectrum. A perfect match between the simulated spectrum and a reference spectrum yields a score of zero.

Because the ability to simulate the shifts for side-chain atoms using the existing model library was lacking, new models were developed. A data set of 153 unique side-chain atoms was compiled using 10 cyclopentanols, 6 cyclopentanes, 4 norbornanols, 3 norbornanes, 5 cyclohexanones, 14 cycloheptanones, 1 hydroxysteroid, 2 ketosteroids, 9 benzenes, 8 PAHs, 2 fluorinated PAHs, and 4 phenols. The side chains represented included the following: ethyl, *n*-propyl, isopropyl, *sec*-butyl, isobutyl, *tert*-butyl, and neopentyl. Although a majority of these compounds have been used in prior studies, models were not developed specifically for side-chain atoms because there were not enough of them in each individual study to generate statistically valid models. The study involving cyclopentanes and cyclopentanols⁹ was the only previous study which included side-chain atoms in the development of regression models. The models in the cyclopentane and cyclopentanol study, however, were not specifically developed for side-chain atoms. Compounds I–IV (Figure 3) were not used in the development of the regression models but were held aside as an external prediction set.

This is our first study combining atoms from a diverse set of structural environments. In all previous studies, all the atoms used in the development of regression models originated

Table V. Statistics for Side-Chain Models

atom subset	<i>N</i>	<i>d</i>	<i>R</i>	<i>s</i> , ppm
primaries	75	5	0.989	1.05
secondaries	24	4	0.993	1.27
tertiaries	19	3	0.982	0.88
quaternaries	35	6	0.968	0.44

from compounds of the same class. This study, however, included atoms contained in compounds which varied in size and functionality. This was also the first study in which the goal was to improve our ability to simulate the shifts for atom types (side-chain atoms on cyclic compounds). The aim of all previous studies was to improve our ability to simulate the spectra for specific compound types (e.g., hydroxysteroids).

In order to develop high-quality models, the data set was divided into four smaller, more homogeneous atom subsets based on connectivity. The 153 atoms were divided into 75 primary, 24 secondary, 19 tertiary, and 35 quaternary carbons. Each subset contained atoms connected to both aromatic and aliphatic ring systems. Using the methodology outlined above, regression models were developed independently for each atom subset. Table V lists the statistics for each model, where *N* denotes the number of observations, *d* is the number of descriptors used, *R* is the coefficient of multiple correlation, and *s* is the standard error of regression. The standard errors range from a low of 0.44 ppm to a high of 1.27 ppm and are in a range that is acceptable. These four models were included in a new, expanded regression model library as models 72–75.

The spectra for the same four test compounds, compounds I–IV, were simulated using the nearest neighbor model selection algorithm to search the updated library containing the newly developed side-chain models. In all cases, side-chain models were automatically selected. As shown in Table III, Section B, the mean residual error for the side-chain predictions using the selected models was 0.51 ppm, a tremendous improvement. The updated library also reduced the average nearest neighbor distance for the eight side-chain atoms from 1.15 to 0.16, which can be interpreted as a measure of the confidence in the predictions. In addition, the overall standard errors of estimate, *s*, significantly improved for each compound.

Library searching was performed to measure further the accuracy of the simulated spectra for these compounds when generated using the new side-chain models. As shown in Table IV, Section B, the actual, authentic spectrum was retrieved as the number one match for each simulated spectrum. More impressively, the scores for each match were all small, thus indicating a high degree of similarity between the simulated spectra and their known-reference spectra. Even though there are only two unique side-chain atoms in each compound, the ability to simulate their shifts accurately using the newly

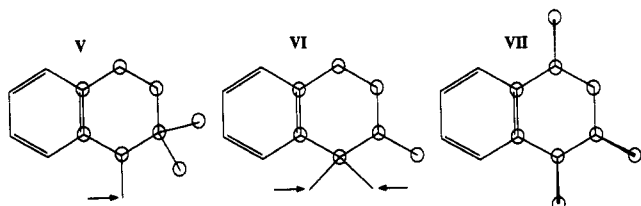


Figure 4. Hydroaromatic compounds used to probe the regression model library.

Table VI. Nearest Neighbor Prediction Results for Probe Compounds V–VII

	before ^a		after ^b	
	mean residual error, ppm	mean nearest neighbor distance	mean residual error, ppm	mean nearest neighbor distance
12 nonbridging aromatic carbons	1.48	0.456	1.48	0.456
24 circled carbons	17.72	1.012	1.12	0.025

^a Before hydroaromatic models were developed and present in regression model library. ^b After hydroaromatic models were developed and available in regression model library. All 24 circled carbons in compounds V–VII selected the newly developed hydroaromatic models.

developed models tremendously improved the overall accuracy of the entire spectra.

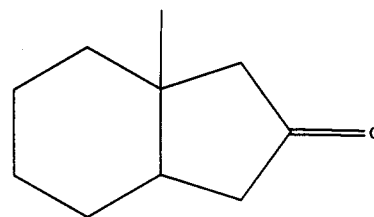
HYDROAROMATIC COMPOUNDS

The ability to simulate the ¹³C NMR spectra for hydroaromatic compounds using the existing model library was tested using compounds V–VII shown in Figure 4. Using the nearest neighbor model selection algorithm, models were selected to predict the shifts for each unique carbon in compounds V–VII. The results for these predictions are summarized in Table VI. The mean residual error for the 12 nonbridging aromatic carbons was 1.48 ppm. These predictions were very accurate because models 54 and 56, models developed for benzenes and PAHs, were selected by the nearest neighbor procedure for the simulation of these shifts. The observed shifts for the three methyl carbons marked with arrows in compounds V and VI of Figure 4 were unavailable. Therefore, the accuracy of the predicted shifts for these atoms could not be determined. The mean residual error for the 24 circled bridgehead and saturated ring carbons in compounds V–VII was 17.72 ppm. These predictions were poor because models have not yet been developed for bridgehead carbons or saturated carbons in hydroaromatic ring systems. The nearest neighbor distances for the 24 circled atoms have a mean greater than 1.0; this large value indicates that the query atoms and the selected nearest neighbor atoms are in significantly different surroundings. Since the ability to simulate accurate shifts for these atoms using the existing model library was lacking, new models were developed which were capable of simulating their shifts.

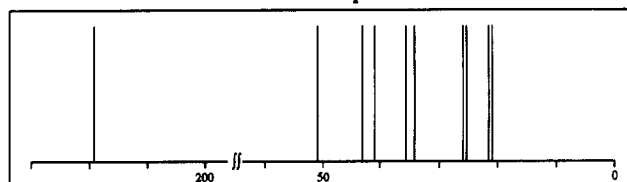
A set of 64 hydroaromatic compounds and their known ¹³C NMR spectra comprised the data set.²² A total of 417 unique bridgehead and saturated carbons and their corresponding chemical shifts were used to develop the models. This set of carbons was divided into four atom subsets: 122 bridgehead carbons; 113 carbons one bond from the bridgehead; 88 primary carbons two or more bonds from the bridgehead; and 94 secondary, tertiary, and quaternary carbons two or more bonds from the bridgehead. Utilizing the previously outlined methodology, models were developed for each atom subset.

Table VII. Statistics for Hydroaromatic Models

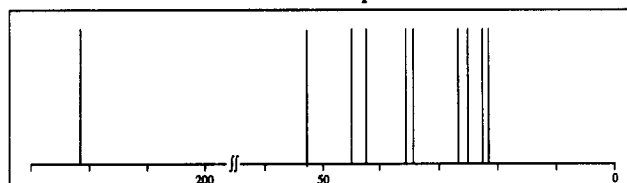
atom subset	N	d	R	s, ppm
bridgehead carbons	122	2	0.960	1.09
carbons one bond from bridgehead	113	8	0.987	0.80
1° ≥ 2 bonds from bridgehead	88	5	0.946	1.75
2°, 3°, & 4° ≥ 2 bonds from bridgehead	94	8	0.991	0.97



Observed Spectrum



Simulated Spectrum



s = 1.69 ppm

Figure 5. Simulated versus observed spectra for *cis*-8-methylhydrindan-2-one using manually selected models.

The statistics for the generated models are listed in Table VII. The standard errors are in the range of 0.80–1.75 ppm. These four models were checked into the library of regression models as models 76–79.

With the new hydroaromatic models present in the library, the nearest neighbor model selection algorithm was utilized to determine if the new models would be selected and if the predictions would improve for compounds V–VII. The newly developed hydroaromatic models were chosen for all the bridgehead and saturated carbons in compounds V–VII. As indicated in Table VI, the predictions for the 24 circled carbons in compounds V–VII improved significantly when the hydroaromatic models were selected from the regression model library. The mean residual error improved from 17.72 ppm before the models for hydroaromatic compounds were present to 1.12 ppm after they were present in the regression model library. In addition, the mean nearest neighbor distance improved from 1.012 to 0.025, which indicates greater confidence in the predictions.

As shown above, the nearest neighbor model selection method is extremely effective for selecting appropriate regression models for compounds that are very similar to those used to develop the library models. Furthermore, if the selected models are statistically robust, accurate predictions can be obtained. However, the current nearest neighbor approach is not always successful for compounds that are not similar to the ones used to develop the library models. For example, the nearest neighbor approach was unable to select models which could accurately simulate the spectra for *cis*-8-methylhydrindan-2-one shown in Figure 5. Although models have

never been specifically developed for hydrindanones, models have been developed for similar compounds such as cyclopentanones, decalones, and ketosteroids. As shown in Figure 5, when models developed from these similar compound classes are manually selected from the regression model library, the spectrum for the molecule can be simulated with an RMS error less than 2 ppm. This example points to a second weakness with the current approach to model selection. The current procedure (based on nearest neighbor selection using seven specially designed environmental descriptors) cannot always select appropriate models from the database when they exist. This may be due to the matching criterion used, or it may be due to the selection of environmental descriptors being used. In addition to developing new models to expand the breadth of the model library, work is concurrently being done to improve the model selection procedure.

In addition to the work described above, the applicability of artificial neural networks as a supplement to linear regression analysis in relating calculated atom-centered descriptors to chemical shifts is also being studied. A set of previously examined ketosteroid molecules has been studied using neural networks, and the results were compared to regression analysis.²³

CONCLUSIONS

It has been shown that the nearest neighbor model selection algorithm can be used systematically to determine weaknesses in the regression model library. Using calculated atomic descriptors and multiple linear regression analysis, high-quality mathematical models can be developed which address such weaknesses. This approach was illustrated for side-chain atoms on cyclic compounds and for selected atoms in hydroaromatic compounds, and a tremendous improvement in our ability to simulate the ¹³C NMR shifts for these carbon environments was reported.

This new model selection procedure has the broad applicability to be useful in the pursuit of the overall goal of this research, which is the development of an automated system capable of simulating the ¹³C NMR spectra of organic compounds.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation Grant CHE-8815785 and by the Department of Defense Grant DAAL03-89-G-0069. The Sun 4/110 workstation was purchased with partial financial support of the National Science Foundation. Portions of this paper were presented at the 202nd National Meeting of the American Chemical Society, New York, NY, Aug 1991.

REFERENCES AND NOTES

- (1) Zupan, J.; Novič, M.; Bohanec, S.; Razinger, M.; Lah, L.; Tušar, M.; Košir, I. Expert System for Solving Problems in Carbon-13 Nuclear Magnetic Resonance Spectroscopy. *Anal. Chim. Acta* **1987**, *200*, 333–345.
- (2) Fürst, A.; Pretsch, E. A Computer Program for the Prediction of ¹³C-NMR Chemical Shifts of Organic Compounds. *Anal. Chim. Acta* **1990**, *229*, 17–25.
- (3) Bremser, W. Hose—A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355–365.
- (4) Yuan, S.-G.; Wang, Y.-W.; Chen, D.; Zheng, C.-Z. A Computer Search System for Similar Organic Compounds in Carbon-13 Nuclear Magnetic Resonance Data Files. *Anal. Chim. Acta* **1989**, *221*, 345–351.
- (5) Lindeman, L. P.; Adams, J. Q. Carbon-13 Nuclear Magnetic Resonance Spectrometry. *Anal. Chem.* **1971**, *43*, 1245–1252.
- (6) Smith, D. H.; Jurs, P. C. Prediction of ¹³C Nuclear Magnetic Resonance Chemical Shifts. *J. Am. Chem. Soc.* **1978**, *100*, 3316–3321.
- (7) Small, G. W.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Cycloalkanols with Computer-Based Structural Descriptors. *Anal. Chem.* **1983**, *55*, 1128–1134.
- (8) Small, G. W.; Jurs, P. C. Data Reduction in the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Steroids. *Anal. Chem.* **1984**, *56*, 2307–2314.
- (9) Egolf, D. S.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Substituted Cyclopentanes and Cyclopentanols. *Anal. Chem.* **1987**, *59*, 1586–1593.
- (10) Egolf, D. S.; Brockett, E. B.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Methyl-Substituted Norbornan-2-ols. *Anal. Chem.* **1988**, *60*, 2700–2706.
- (11) Sutton, G. P.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Alkyl-Substituted Cyclohexanones and Decalones. *Anal. Chem.* **1989**, *61*, 863–871.
- (12) Ranc, M. L.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Piperidines. *Anal. Chem.* **1989**, *61*, 2489–2496.
- (13) Egolf, D. S.; Jurs, P. C. Structural Analysis of Polychlorinated Biphenyls from Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1990**, *62*, 1746–1754.
- (14) Sutton, G. P.; Jurs, P. C. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Alkyl-Substituted Aromatic Compounds. *Anal. Chem.* **1990**, *62*, 1884–1891.
- (15) Sutton, G. P.; Anker, L. S.; Jurs, P. C. Evaluation of Automated Methods for the Selection of Models for Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Keto-Steroids. *Anal. Chem.* **1991**, *63*, 443–449.
- (16) 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.
- (17) Ball, J. W.; Anker, L. S.; Jurs, P. C. Automated Model Selection for the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Cyclopentanones and Cycloheptanones. *Anal. Chem.* **1991**, *63*, 2435–2442.
- (18) Jurs, P. C.; Sutton, G. P.; Ranc, M. L. Carbon-13 NMR Spectral Simulation. *Anal. Chem.* **1989**, *61*, 1115A–1122A.
- (19) Small, G. W.; Jurs, P. C. Determination of Topological Similarity of Carbon Atoms in the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1984**, *56*, 1314–1323.
- (20) Small, G. W.; Stouch, T. R.; Jurs, P. C. Automated Selection of Models for the Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra. *Anal. Chem.* **1984**, *56*, 2314–2319.
- (21) Randić, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164–175.
- (22) *Conformational Analysis of Cyclohexenes, Cyclohexadienes, and Related Hydroaromatic Compounds*; Rabideau, P. W., Ed.; VCH: New York, 1989; Chapter 5.
- (23) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157–1164.