Automated Spectrum Simulation Methods for Carbon-13 Nuclear Magnetic Resonance Spectroscopy Based on Database Retrieval and Model-Building Strategies

Robert C. Schweitzer and Gary W. Small*

Center for Intelligent Chemical Instrumentation, Department of Chemistry, Ohio University, Athens, Ohio 45701-2979

Received December 18, 1996[⊗]

An automated method for the prediction of carbon-13 nuclear magnetic resonance (^{13}C NMR) chemical shifts which involves the combined use of database retrieval and multivariate calibration methods is described. A commercial database of chemical structures and experimentally measured and assigned ^{13}C NMR spectra is used to implement this methodology. The chemical environment of each of the carbons in the database is encoded as a vector, and Euclidean distance comparisons are made between these vectors and the similarly encoded environment of a target carbon whose chemical shift is to be predicted. This procedure yields a calibration set of carbon atoms with chemical environments similar to that of the target carbon. If the environment of the closest match is very similar to that of the target carbon, the predicted chemical shift is assigned as the chemical shift of that closest match (i.e., a direct chemical shift retrieval is performed). Otherwise, a chemical shift model based on spectra−structure relationships is built using the selected calibration set of carbons. The predicted chemical shift is then calculated with the computed model. The independent variables in the model are derived from numerical structural descriptors which describe some topological, geometrical, or electronic aspect of the chemical environment of the carbon atoms in the calibration set. Multiple linear regression, partial least-squares regression, and principal component regression are compared in terms of their effectiveness for use in building the chemical shift models. A series of experiments is performed to test the accuracy of direct chemical shift retrieval versus model building and to determine the optimal settings of several parameters that affect the model-building step. Based on a test set of 38 504 carbons, an overall mean deviation of 1.85 ppm between predicted and actual chemical shifts is achieved by use of direct retrieval alone, while the corresponding mean deviation based on a combination of direct retrieval and model building is 1.69 ppm.

INTRODUCTION

Carbon-13 nuclear magnetic resonance spectroscopy (^{13}C NMR) is a technique which has found wide use in the solution of organic structure elucidation problems. Structural changes in the carbon skeleton of a molecule can be readily detected by ^{13}C NMR, and this sensitivity produces ^{13}C NMR spectra which are rich in information and are challenging to interpret. Consequently, interest has grown in the development of computer-based techniques to aid in spectral interpretation. One category of these computer-assisted structural elucidation tools is spectrum simulation.

In practice, ^{13}C NMR spectrum simulation is based on predicting the chemical shift of each carbon atom in a given structure. The resulting simulated spectrum can be compared with experimentally measured spectra to test a hypothesis about an unknown structure. Two practical approaches for estimating the chemical shift of a carbon atom are direct database retrieval methods[1−6] and empirical modeling techniques.[7−12]

The direct retrieval approach makes use of a database of chemical structures and their corresponding experimentally measured and assigned ^{13}C NMR spectra. Every carbon atom in each structure is represented algorithmically in some manner that relates its chemical environment to its chemical shift. The prediction of a chemical shift for a target carbon is then performed by searching the database of known chemical environments to find the closest match to the target chemical environment. The chemical shift of the closest match is then assigned as the predicted shift of the target carbon. The use of this procedure for each carbon atom in a given test structure results in a complete predicted ^{13}C NMR spectrum. The advantage of the direct retrieval approach is that chemical shift predictions can be performed for a wide variety of chemical environments. The disadvantage is that an accurate predicted shift requires the presence in the database of a carbon with a chemical environment highly similar to the chemical environment of the target carbon.

In the empirical modeling method, models are computed that relate chemical structural features to observed chemical shifts. The form of the models typically used is

$$s_a = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \qquad (1)$$

where s_a is the chemical shift of atom a, the b_i terms are weighting coefficients, and the x_i terms are either calculated structural descriptors that encode some topological, electronic, or geometrical aspect of the chemical environment of atom a or some linear combination of these descriptors. A set of atoms with known chemical shifts is used to calculate the b_i terms through the use of statistical model-building techniques such as multiple linear regression (MLR), partial least-squares regression (PLSR), or principal component regression (PCR). To predict a ^{13}C NMR spectrum

* Author to whom correspondence should be sent. Tel: (614) 593-1748. E-mail: small@ohiou.edu.

[⊗] Abstract published in *Advance ACS Abstracts,* February 15, 1997.

S0095-2338(96)00173-4 CCC: $14.00

for a given structure, previously developed models are used to calculate the chemical shifts of each of the carbons in the structure and the resulting shifts are assembled into a spectrum. The empirical modeling method can often predict chemical shifts very accurately, but a given model is only able to operate effectively when it can interpolate among the environments and chemical shifts used to build the model. A significant advantage brought about by the interpolative nature of the method, however, is that an exact match to the target carbon does not have to occur among the carbons used to construct the model.

This paper describes efforts to combine the two approaches, thereby obtaining the generality of the direct retrieval method and the interpolative capability of the model-building approach. The first step in this combined method is the use of an encoding algorithm to create an atom-based database in which each of the carbon atoms is encoded according to its chemical environment and stored along with its experimentally observed chemical shift. A small subset of carbon atoms in the database whose chemical environments are most similar to the chemical environment of a target carbon in a test structure is retrieved by use of the direct retrieval approach. If the match in environments is close enough, the predicted chemical shift can be assigned as the experimental chemical shift associated with the closest matching environment. Otherwise, this set of retrieved carbon atoms with known chemical shifts is used to build an empirical model, and the resulting model is used to calculate the shift of the target carbon. A method is developed in which the required models are built in a completely automated manner in real time. This allows each model to be tailored to the target carbon to the maximum degree possible. The procedure is repeated for every carbon in the test structure, and the resulting predicted chemical shifts can be used to assemble a complete simulated spectrum.

EXPERIMENTAL SECTION

The Sadtler ^{13}C NMR database consisting of 29 966 structures and their accompanying experimentally measured and assigned chemical shifts was used for this research (Bio-Rad, Inc., Sadtler Division, Philadelphia, PA). The structures consisted of organic compounds ranging in size from 1 to 40 carbons and contained a large number of different functional groups. The use of geometrical-based descriptors required the conversion of the structures from two-dimensional to three-dimensional representations. This was performed by use of a molecular mechanics procedure reported by Stuper et al.[13] The coordinates obtained through this calculation were then refined by use of MM2 (1987 version), a molecular mechanics package developed by Allinger and co-workers.[14] A set of 21 199 structures was chosen from the original 29 966 structures for the work reported in this paper. The data set was divided randomly into two sets of structures. The larger set contained 16 959 structures and was used to define the database (library set) used subsequently in the chemical shift retrieval experiments. The smaller set, which contained 4240 structures, was used as a test set. A random subset of 423 of the 4240 structures was used as a test subset for some of the experiments to reduce the required computational time. The best results were confirmed by use of the full test set.

The spectrum simulation computations were performed on a Silicon Graphics 4D/460 computer system running under Irix (version 5.2; Silicon Graphics, Mountain View, CA) and operating in the Center for Intelligent Chemical Instrumentation at Ohio University. The software developed for this work was written in FORTRAN 77. MLR routines from the IMSL library (IMSL, Inc., Houston, TX) were used in the model-building calculations. Analysis of variance (ANOVA) calculations were performed with the Minitab software package (version 10.5; Minitab, Inc., State College, PA) implemented on a Dell 466/L computer operating under MS-DOS (version 6.2; Microsoft, Inc., Redmonnd, WA) and Microsoft Windows 3.1.

RESULTS AND DISCUSSION

Environmental Encoding Algorithm. The first step in the spectrum simulation algorithm is the use of the direct retrieval method to select a subset of carbon atoms from the database which have similar chemical environments to the target carbon whose predicted chemical shift is sought. An algorithm is thus required to encode the chemical environments of each carbon in the database and the carbons in structures input for shift prediction in a manner that correlates with chemical shifts and that allows rapid and quantitative comparisons. The algorithm used in this work was originally developed by Small and Jurs[15] and recently extended and improved by Schweitzer and Small.[16] In this scheme, the environment of carbon atom a is represented as an $n + 1$ dimensional vector of the form

$$\mathbf{e}_a = (e_0, e_1, ..., e_n) \tag{2}$$

where the first dimension, e_0, characterizes atom a, the second dimension, e_1, describes the influence of atoms one bond removed from atom a on its chemical shift, and each successive dimension describes the collective influence of the atoms one bond further from atom a on the chemical shift of that atom. The set of parameters which are used to compute the elements of $\mathbf{e}_a$ was developed on the basis of experimental chemical shifts of small molecules. In the work reported here, n in eq 2 was set to 7 and a double vector was used as described in ref 16, resulting in a 13-dimensional vector.

A subset of carbons from the database is retrieved by comparing the vector generated for the target carbon to each of the vectors representing carbons in the database. A Euclidean distance calculation is used to perform these vector comparisons. For the work presented here, the 200 carbons whose vectors gave the smallest Euclidean distances were included in the retrieved subset. When model building is required, the selected subset of carbons is used in the model-building step. It should be noted that the environmental encoding algorithm is used only to find the subset of carbons and has no further use in the model-building step.

Combining Direct Retrieval and Model Building. It has been found that for cases in which the best match has a very small Euclidean distance (score), the chemical shift is best predicted by direct retrieval (i.e., using the shift of the closest match as the predicted shift) rather than by modeling. This section explores the relationship between the accuracy of the predicted shift and the conditions required for direct retrieval to be used.

Table 1. Comparison of Score Cutoffs for Direct Retrieval Experiments

score threshold	no. of carbons	mean deviation (ppm)	max absolute residual (ppm)	median absolute residual (ppm)
1×10^{-7}	700	0.45	11.90	0.10
1×10^{-6}	805	0.44	11.90	0.10
1×10^{-5}	1023	0.44	11.90	0.10
1×10^{-4}	1206	0.48	11.90	0.10
1×10^{-3}	1550	0.56	20.00	0.20
0.01	2061	0.65	20.00	0.20
0.1	2730	0.84	20.00	0.30
1.0	3387	1.17	35.00	0.40
10	3760	1.59	52.80	0.40
none	3849	1.80	53.98	0.50

The experiments reported were performed for the smaller test set of 423 structures (3849 carbons). Table 1 displays the relationship between the score and the accuracy of the predicted chemical shift. Column 1 contains the score cutoff value for the use of direct retrieval for shift prediction. For each cutoff, shift prediction by direct retrieval was performed if the Euclidean distance between the environment vectors of the target carbons and the closest matching carbon in the database was less than the cutoff. Column 2 lists the number of carbons in the test set which met this criterion. The mean deviation between predicted and actual chemical shifts is listed in the third column, while columns 4 and 5 list the maximum and median absolute residuals, respectively, for the chemical shifts predicted by direct retrieval. It can be seen from this table that the smaller the score cutoff, the better the predicted shift but the fewer the number of shifts predicted by direct retrieval. A study was conducted to find the percentage of predicted shifts with absolute residuals greater than some tolerance for a series of score cutoffs, and that percentage was nearly constant for score cutoffs from 1×10^{-4} to 0.01. The percentage was observed to increase dramatically, however, for score values from 0.01 to 10.0. For example, the percentages of predicted chemical shifts with absolute residuals greater than or equal to 3.0 ppm for score cutoffs of 0.01, 0.1, 1.0, and 10.0 were 5.2%, 7.4%, 11.6%, and 15.7%, respectively.

Further restrictions on the use of direct retrieval have been studied to obtain more accurately predicted shifts. One such condition is that the standard deviation of the retrieved chemical shifts for the top *n* matches be smaller than some cutoff value. The rationale for using this condition is that if the retrieved shifts are all similar and thus the standard deviation is small, similar chemical environments have been found, and it is likely that the chemical shift of the closest matching environment will be an accurate estimate of the desired predicted shift. The value of the standard deviation cutoff was varied from 0.1 to 20 ppm for score cutoffs ranging from 1×10^{-7} to 10.0. If the standard deviation cutoff is stringent, very few carbons are predicted by direct retrieval, but if the cutoff is lenient, no improvements in the accuracy of the predictions are seen. Table 2 shows the result of varying the standard deviation cutoff from 0.1 ppm to no cutoff for a score cutoff of 1.0. The first column lists the standard deviation cutoff. Columns 2−5 have the same headings as for Table 1. It can be seen that as the standard deviation increases to around 5.0 ppm and beyond, very few of the carbons are excluded from prediction by direct retrieval. A value between 2.0 and 5.0 ppm seems to provide a good balance between the accuracy of the predicted chemical shift and the problem of excluding too many carbons from being predicted by direct retrieval.

Table 2. Evaluation of Standard Deviation Cutoff for Chemical Shift Retrieval

cutoff (ppm)	no. of carbons	mean deviation (ppm)	max absolute residual (ppm)	median absolute residual (ppm)
0.1	83	0.07	0.30	0.10
0.5	809	0.29	12.40	0.10
1.0	1460	0.47	12.40	0.20
1.5	1942	0.60	12.40	0.20
2.0	2252	0.71	12.40	0.30
3.0	2696	0.90	20.00	0.30
4.0	2982	1.01	20.00	0.30
5.0	3132	1.08	20.00	0.40
10.0	3330	1.16	35.00	0.40
15.0	3358	1.17	35.00	0.40
20.0	3379	1.17	35.00	0.40

Structural Descriptors. The use of model building requires a set of structural descriptors from which the independent variables for the model will be selected. These descriptors encode some aspect of the chemical environment of the target carbon whose predicted chemical shift is sought. Several hundred structural descriptors have been reported in previous work devoted to the development of empirical chemical shift models.[7−12] From these descriptors, 121 were selected as representative for the work presented in this paper. Three general classes of descriptors were used as follows: (1) topological descriptors, (2) geometrical descriptors, and (3) electronic descriptors. Topological descriptors are based on the topology of the molecule and include such calculations as the number of secondary sp^3 carbons three bonds from the target carbon or the number of Cl atoms four bonds from the target carbon. Geometrical descriptors are based on the modeled three-dimensional atomic coordinates of the structures and include calculations such as van der Waals energies or inverse throughspace distances between the target carbon and other structural features in the molecule. Electronic descriptors attempt to describe the electronic structure of the target carbon and its surrounding environment. These descriptors are based on atomic and molecular orbital calculations.

A total of 64 topological descriptors in five classes were used. The α valency count descriptor (AVC) counts the number of non-hydrogen atoms of a particular valency (1°, 2°, 3°, or 4°) *n* bonds (1 or 2) from the target carbon. The nearest neighbor descriptor (NNA) counts the number of atoms of a particular type (O, N, F, Cl, Br, I, S, or P) *n* bonds (1, 2, 3, or 4) from the target carbon. The bond type count descriptor (BVC) counts the number of atoms involved in a particular bond hybridization (double, triple, or aromatic) *n* bonds (1, 2, 3, or 4) from the target carbon. The functional group count descriptor (FGD) counts the number of atoms involved in a particular functional group (carboxylic carbons, ketone carbons, amide carbons, hydroxyl oxygens, or nitro group nitrogens) *n* bonds (1 or 2) from the target carbon. The valency-corrected connectivity index descriptor (CCI) for atoms *n* bonds (1 or 2) from the target carbon accounts for the two remaining topological descriptors and functions by describing the degree of branching around the target carbon.[17]

The set of geometrical descriptors consists of 32 descriptors in five classes. These descriptors encode the steric environment of the target carbon and are primarily based

on computed throughspace distances from the target carbon to other atoms in the molecule. The heteroatom distance descriptors (HAD) sum the inverse cubed throughspace distances from the target carbon to eight different heteroatoms (O, N, F, Cl, Br, I, S, and P). The hydrogen radial distance descriptor (HRD) calculates the sum of the inverse cubed throughspace distances to hydrogens *n* bonds (2, 3, 4, or 5) from the target carbon. The radial distance descriptor (RDD) calculates the sum of the inverse cubed throughspace distances to all non-hydrogen atoms *n* bonds (1, 2, 3, or 4) from the target carbon. The step α hydrogen descriptors (SAH) consist of three different types. The first and second sets of SAH descriptors calculate the sum of the inverse cubed throughspace distances for all hydrogens attached to the target carbon to the hydrogens *n* bonds (2, 3, 4, or 5) from the target carbon and to the non-hydrogen atoms *n* bonds (2, 3, 4, or 5) from the target carbon. The third set of SAH descriptors calculates the sum of the inverse cubed throughspace distances for all hydrogens attached to the target carbon to the hydrogens three to five bonds inclusively from the target carbon. The van der Waals energy descriptors (VDW) encode nonbonded interactions in the structure and are computed with the same energy function used in Allinger's MM2 molecular mechanics program.[14] Six such descriptors are included. The first is the van der Waals energy due to all interactions between the hydrogens attached to the target carbon and all hydrogens from two to seven bonds from the target carbon. The second is the van der Waals energy due to all interactions between the hydrogens attached to the target carbon and all non-hydrogen atoms from two to seven bonds from the target carbon. A third VDW descriptor is computed as the sum of the first two VDW calculations. The fourth through sixth descriptors are the same as the first through third with the exception that the VDW interactions are based on the target carbon itself and not its attached hydrogens.

There are two classes of electronic descriptors, and these account for a total of 25 descriptors. The first class contains electronic descriptors (ELD) based on a simple Hückel calculation.[18−22] The ELD descriptor used in this work is superdelocalizability, which is a measure of chemical reactivity. The first of the 13 ELD descriptors calculates the superdelocalizability of the target carbon. The remaining ELD descriptors calculate the average, minimum, and maximum superdelocalizabilities of all atoms *n* bonds (1, 2, 3, or 4) from the target carbon. The second electronic descriptor is the step charge descriptor (SCG) and is based on a σ charge calculation.[23] The first of the 12 SCG descriptors calculates the σ charge on the target carbon. The next five descriptors calculate the average σ charge for all atoms *n* bonds (1, 2, 3, 4, or 5) from the target carbon. The remaining six descriptors calculate the most positive and most negative σ charges among all atoms *n* bonds (1, 2, or 3) from the target carbon.

Exploration of Model Building. Equation 1 describes the model-building step in a concise form. The first step in the procedure, the selection of the set of carbons with known chemical shifts, has been described previously. The second step is the use of a model-building technique with the calibration set of calculated structural descriptors and known chemical shifts. This work compared MLR, PLSR, and PCR for use in implementing the model-building step.[24−26]

Previous work in our laboratory directed to the calculation of chemical shift models has been performed interactively and employed MLR exclusively. In this procedure, the model developer runs experiments with different combinations of structural descriptors in an attempt to find the best possible descriptor subset for use in building the chemical shift model. The pool of possible descriptors is screened to eliminate descriptors that contain little information (e.g., those with small relative standard deviations or those that are highly correlated with other descriptors). Then, chemical shift models are constructed through the use of variants of MLR that include subset selection (e.g., stepwise MLR).

In the work described here, it was necessary to automate this process. The 121 descriptors defined the starting pool of descriptors. A subset of this set of descriptors was selected for use in the model building through a series of eliminations. Descriptors in which a large percentage of the values were zero or in which the relative standard deviation of the values was very small were eliminated from consideration. A pool of descriptors was then chosen for use with stepwise MLR. The descriptor exhibiting the maximum correlation coefficient with the dependent variable of chemical shifts was selected as the first descriptor in the pool. The remaining descriptors for the pool, up to some maximum number, were chosen one at a time by a stepwise orthogonalization procedure. This calculation treats the set of values comprising a descriptor as the coordinates of a multidimensional vector. Gram−Schmidt orthogonalization[27] was used with the *p* chosen descriptor vectors to compute a set of *p* orthonormal basis vectors that described the data space spanned by the descriptors. The next descriptor added to the pool corresponded to that descriptor vector which was most orthogonal to the current set of basis vectors (i.e., the descriptor that added the most unique information to the *p* descriptors already chosen). The selected descriptor was then used to calculate the $p + 1$ basis vector that described the updated data space now spanned by the $p + 1$ descriptor vectors. This process was repeated until the descriptor pool reached the specified size.

The pool of descriptors that resulted from this procedure was submitted to stepwise MLR for the calculation of the chemical shift model. Stepwise MLR progressively builds the model from 1 to *n* terms by adding at each step the descriptor that explains the greatest fraction of remaining variance in the dependent variable of chemical shifts. The first variable entered is simply that descriptor that correlates most strongly with the dependent variable. The entry of a variable into the model at each step is controlled by a statistical *F*-test at a user-specified probability.

The two other regression techniques studied in this work were stepwise PLSR and stepwise PCR. Both techniques use the structural descriptors as a set of input variables and compute a new set of latent variables based on linear combinations of the inputs. The criteria used in the construction of the latent variables serve to extract orthogonal components from the descriptor pool. The independent variables used in the model-building step are these latent variables, termed partial least-squares (PLS) factors or principal components (PCs), respectively, in PLSR and PCR. In the implementation used here, a specified maximum number of PLS factors or PCs was calculated and the independent variables for the model were then selected from this pool by use of stepwise MLR.

Table 3. Parameter Settings for Model-Building Experiments

variable	levels
limit for use of direct retrieval (dr)	0.01, 0.1, 1.0
limit for inclusion of carbons in calibration set (ml)	5.0, 20.0, 10000
entrance criterion for stepwise regression (ec)	0.001, 0.01
maximum number of factors in model (mf)	5, 10, 15
maximum number of descriptors in pool (md)	20, 40, 60
percentage of zero values cutoff (zc)	50%, 75%, 95%
relative standard deviation cutoff (rsd)	5%, 15%, 25%

An advantage of methods based on the construction of latent variables is that they have the capability to extract relevant information from the descriptor pool without the necessity for screening or subsetting the descriptors prior to the calculation. This allows greater automation of the model-building procedure.

Investigation of Model-Building Parameters. A study was performed to investigate the parameter settings used with MLR, PCR, and PLSR. These parameters are listed in Table 3 along with the levels used. The first parameter, dr, is the score cutoff used to determine whether direct retrieval or model building is used for the chemical shift prediction. Direct retrieval is used if the Euclidean distance of the closest matching environment vector is less than the specified dr value.

The second parameter studied was denoted ml and is the maximum allowed value of the score for a carbon to be included in the calibration set used in building the model. If the score is larger than this cutoff, it is assumed that the chemical environment of that carbon is too different from the environment of the target carbon. A related issue is the fact that the number of observations in the calibration set must be greater than the number of independent variables employed in the regression analysis. For all three regression techniques, a minimum of 30 observations was required for model building to be used for chemical shift prediction. If this criterion was not met, direct retrieval was used. Thus, if the score requirement for carbons to be included in the calibration set is very stringent, fewer calibration sets will have the required minimum of 30 observations and fewer shifts for carbons in the test set will be predicted by models than if the score requirement for entry into the calibration set were less stringent. The least restrictive level for this entrance criterion was set at 10 000 to ensure that all 200 carbons retrieved were included in the calibration set.

The ec parameter denotes the entrance criterion for the inclusion of independent variables in the stepwise regression model. It is specified as a probability level for the significance of the variable in the model and is based on the use of an *F*-statistic.

A fourth parameter studied was the maximum number of PLS factors or PCs to use as independent variables in the PLSR or PCR models. It was denoted mf and set at a level below the minimum required value of 30 observations (carbons in the calibration set). A related parameter for MLR is the size of the pool of descriptors from which variables are chosen for model building (md). The number of structural descriptors chosen from this pool for inclusion in the model is not controlled directly but is dependent on the entrance criterion level used. However, larger models will tend to result from a larger pool of independent variables. A trap was also instituted so that the number of descriptors allowed into the pool was always less than one-third the number of observations. Thus, if a relatively small number of observations were present in the calibration set, a correspondingly small number of descriptors were used as independent variables, and the likelihood of a model being built on chance correlations was lessened.

The MLR studies included two additional parameters which were not needed for PLSR or PCR and which were mentioned previously. One parameter is the cutoff value for the use of a given structural descriptor based on the maximum percentage of zero values in the set of observations for that descriptor (zc). The second parameter is the minimum percent relative standard deviation of the values in a given descriptor to allow that descriptor to be used (rsd).

A full factorial experimental design was conducted using every possible combination of parameter settings for each of the three regression techniques. The primary objective of the study was to determine if the combination of model building with direct retrieval could produce predicted chemical shift values which were more accurate than those predicted by use of direct retrieval alone. A second objective was to find the set of parameters which would result in the most accurate predicted chemical shifts. The test set used for these exploratory experiments contained 3849 carbons. For a given experiment, based on the parameter settings, some of the 3849 carbons were predicted by use of model building and the remainder of the carbons were predicted by direct retrieval. For each predicted chemical shift, a residual was calculated as the absolute value of the difference between the known (as assigned in the Sadtler database) and predicted chemical shifts. The mean deviation (i.e., mean absolute residual) based on the 3849 carbons in the test set was calculated and employed as a response function for the experimental design study. This response function was used to assess the ability of the various combinations of parameter settings to produce accurate chemical shift predictions and to judge the significance of the parameters. The mean deviation was used in favor of more traditional statistics such as the root-mean-square error (standard error of prediction) as it is more resistant to the influence of large outlying residuals. This was important as no efforts were made in this part of the study to exclude carbons from prediction in cases in which no closely matching environments were found in the database. As a consequence, there were a few large outliers among the residuals.

PLSR Results. Of the 54 experiments performed with PLSR, the lowest values of the mean deviation between predicted and actual chemical shifts cluster around 1.65 ppm, and the highest values are around 1.84 ppm. The mean deviation if all chemical shifts in the test set are predicted by direct retrieval is 1.80 ppm. The level at which direct retrieval is employed has a noticeable effect with a level of dr = 0.01 giving the worst results (1.80 $\pm$ 0.04 ppm) and the level of 0.1 (1.75 $\pm$ 0.07 ppm) and 1.0 (1.76 $\pm$ 0.08 ppm) giving similar results. The numbers in parentheses denote the mean and standard deviation of the response function values for all experiments in which the specified level of dr was used. Within each dr level, the model limit level makes a noticeable difference. The cases in which all possible carbons were used in the calibration set regardless of score (ml = 10 000) yielded the very best results of the 54 experiments. There were six experiments performed at a dr level of 0.1 for which the ml level was 10 000. Among these six, the range in mean deviations was only 0.03 ppm.

The mean deviations for the corresponding six experiments for which the dr level was 1.0 exhibited a range of only 0.01 ppm. The number of PLS factors (mf) and the entrance criterion (ec) were observed to have relatively little effect. Thus, the best values using PLSR were obtained with an ml level of 10 000 and a dr level of 0.1 or 1.0. The results for a dr level of 1.0 with an ml level of 10 000 (1.65 ± 0.01 ppm) were slightly better than the dr level of 0.1 (1.67 ± 0.01 ppm), but the difference was not large enough to be a determining factor. The case in which a dr level of 0.1 was used predicted 1119 of the 3849 carbons by model building, while only 462 of the 3849 carbons were predicted by model building when dr was set to 1.0. Given a reliable estimate of the ability of a computed model to predict chemical shifts accurately, one could use a dr level of 0.1 and simply eliminate those models which are judged not likely to predict well. Direct retrieval could be used instead for those cases. This process would allow more carbons to be predicted by model building and would result in a better overall prediction accuracy. The development of such a model assessment procedure is discussed in the final section of this paper.

An ANOVA calculation[28] was performed for the 54 PLSR experiments. The main effects and the two-way interactions were studied, and the higher order interactions were combined and treated as the error term. This calculation showed all main effects and two-way interactions to be significant at a 95% level or better. The most important effects in order of significance were (1) ml, (2) dr, and (3) a tie between ec and the two-way interaction between ml and dr. These findings correspond to what was found by the empirical study of the PLSR results.

PCR Results. The results for the 54 experiments run with PCR are not as good as those obtained with PLSR. The best values of the mean deviation for PCR cluster around 1.74 ppm, and the worst values are around 2.20 ppm. The trends for the PCR calcuations are not as clear as those for PLSR. The best results based on the direct retrieval parameter were found for a dr level of 1.0. In fact, the only cases in which the combined direct retrieval/model-building approach did better than direct retrieval alone were those in which the dr level was set at 1.0. Of the 18 experiments for which the dr level was 1.0, the trend for the three levels for the model limit (ml) was that a limit of 5.0 gave the worst results (1.82 ± 0.03 ppm). A limit of 20.0 gave better results (1.78 ± 0.01 ppm), and the results for a ml limit of 10 000 were slightly better still (1.77 ± 0.04 ppm). For the best case where dr = 1.0 and ml = 10 000, the number of PCs used was significant. If only five principal components were allowed, the results were worse (1.82 ppm) than in the cases for which mf = 10 (1.76 ppm) or 15 (1.74 ppm).

The results obtained from the ANOVA calculation for the 54 PCR experiments confirmed the empirical observations discussed above. All main effects and two-way interactions except for the interaction between mf and ec were significant at a 93% level or better. The most important effects in order of significance were (1) dr, (2) ml, (3) a tie between mf and the interaction between dr and ml, and (4) the interaction between ml and mf. An important difference from the PLSR experiments is that the maximum number of PCs (mf) was found to be relatively important.

MLR Results. A total of 486 experiments were run with MLR. The number of descriptors allowed in the pool of descriptors (md) had a noticeable effect, with a limit of 60 giving better results than a limit of 20 or 40. This observation seems intuitive since it allows a greater choice in the descriptors used. Consequently, all of the results reported for MLR are solely for the case in which a pool of 60 descriptors was allowed (162 total experiments). In addition, the criterion that a certain minimum percent relative standard deviation in the descriptor values be met had no effect as identical results were obtained for each of the levels of 5%, 15%, and 25%. These limits are apparently too high, and consequently, the values for only one level will be considered. The best value of the mean deviation for the remaining 54 MLR experiments was 1.68 ppm, and the worst value was 10.35 ppm. It can be seen that the values calculated using MLR vary much more widely than do those obtained with PLSR or PCR. The best value had parameter levels of zc = 95%, ml = 10 000, dr = 1.0, and ec = 0.001. Several interesting interactions between the parameters were noted. The zc condition, for example, tends to give the best average results (1.86 ± 0.08 ppm) for a value of 50%. This is the most stringent level for this criterion and ensures that only those descriptors which have 50% or more nonzero values will be used in the model. The 95% level has a much worse mean value (2.44 ± 2.04 ppm) with a great deal of variance, but some of the individual values are better than those obtained for a zc level of 50%. The zc level of 95% allows descriptors to be used which are not useful in explaining chemical shift values, but it also allows some descriptors to be used which are useful and which a zc level of 50% does not allow. If the useless descriptors can be prevented from entering the model by some other criterion, the overall model building will be enhanced by the presence of the additional useful descriptors. The ec condition at a value of 0.001 seems to provide this screening capability. In a similar fashion, an ml level of 5.0 provides the best average results (1.84 ± 0.02 ppm), but an ml level of 10 000 when combined with an ec value of 0.001 gives the best individual values. The dr parameter appears to give the best results for a level of 1.0.

The results of the ANOVA calculations for the 162 MLR experiments with md = 60 showed all main effects and two-way interactions to be significant at a >98% level except for the rsd main effect and the two-way interactions involving rsd. The most important effects in order of significance were (1) zc and (2) a tie between the ml and dr main effects.

Comparison of PLS, PCR, and MLR Results. For the small test set of 3849 carbons, PCR exhibited inferior performance compared to PLSR and MLR, whereas PLSR (1.65 ppm) and MLR (1.68 ppm) produced similar mean deviations in chemical shifts. The principal limitation of PCR in this application is that it is an "undirected" latent variable technique. The latent variables are computed simply to model the variance in the set of structural descriptors. Since many of the descriptors are discrete variables (e.g., counts of atoms or bonds) that contain a significant number of zero values, the total variance spanned by the set of descriptors is large, and a small number of PCs does not model a large percentage of the total variance. Thus, the set of PCs does not adequately represent the information content of the full set of descriptors. By comparison, PLSR overcomes this limitation because it is a "directed" latent variable method. The latent variables are computed to explain the covariance between the structural descriptors and the dependent variable of chemical shifts. For this reason,

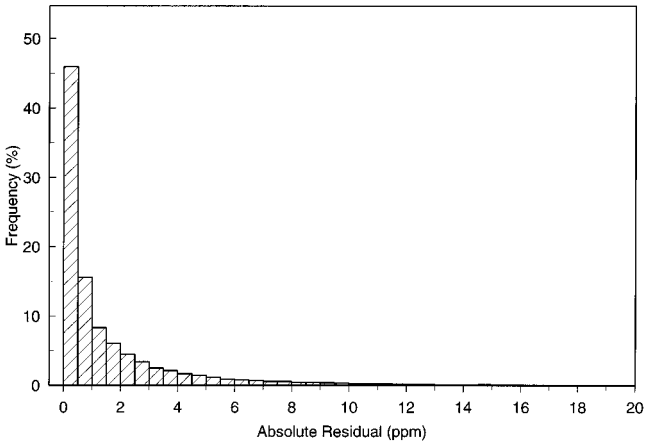

Figure 1. Histogram of the absolute residuals for the 38 504 carbons in the test set for the case in which stepwise PLSR was used with an ml level of 10 000, a dr level of 1.0, an mf level of 10, and an ec level of 0.01.

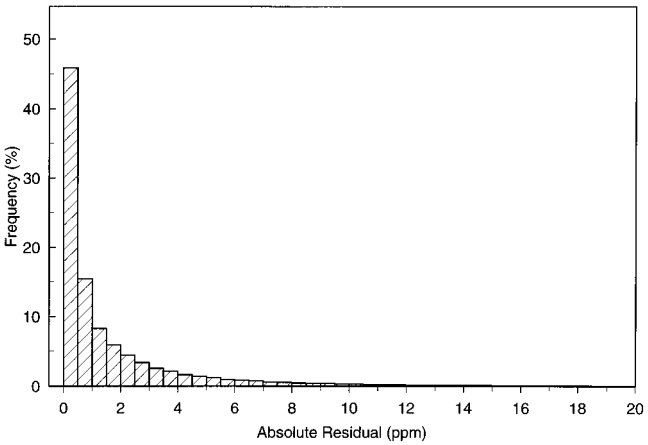

Figure 2. Histogram of the absolute residuals for the 38 504 carbons in the test set for the case in which stepwise MLR was used with an ml level of 10 000, a dr level of 1.0, an md level of 60, an ec level of 0.001, a zc level of 95%, and an rsd level of 15%.

a small number of PLS factors can extract the pertinent information for use in building a chemical shift model.

The full test set of 38 504 carbons was used to test the PLSR and MLR predictive capabilities further for the cases which were determined to be the best with the small test set. The conditions for the best case for PLSR were ml = 10 000, dr = 1.0, mf = 10, and ec = 0.01. The corresponding parameters for the MLR case were ml = 10 000, dr = 1.0, md = 60, ec = 0.001, zc = 95%, and rsd = 15%. The overall mean deviation for the full test set using only direct retrieval was 1.85 ppm. When applied to the large test set, PLSR gave an overall mean deviation of 1.69 ppm with 4382 of the carbons predicted by use of model building. MLR gave an overall mean deviation of 1.70 with 4372 of the carbons predicted using models. Despite the fact that both PLSR and MLR used a setting of dr = 1.0, fewer carbons were predicted with model building for the MLR case due to the occurrence of poorly conditioned data matrices for 10 carbons. The MLR calculation failed to produce a model in these cases. The mean deviation for the MLR calculations was computed without predicted chemical shifts for these carbons.

The set of 38 504 residuals is a distribution of values in which the majority of the residuals are small but a small number of very large residuals exist. Figures 1 and 2 display histograms of the distribution of absolute residuals for the PLSR and MLR experiments, respectively. Both cases have over 60% of the chemical shifts predicted within 1.0 ppm.

The most important question which must be addressed is whether the decrease in the mean deviation from 1.85 to 1.69 or 1.70 ppm is significant. For the set of 4382 carbons whose chemical shifts were predicted by modeling with PLSR, the mean deviation if direct retrieval is used is 6.57 ppm and the mean absolute residual if model building is used is 5.10 ppm. For the set of 4372 carbons whose chemical shifts were predicted by use of MLR, the mean deviation if direct retrieval is used is 6.57 ppm and the mean absolute residual if model building is used is 5.20 ppm. These statistics show more clearly the advantage of using model building for cases in which close chemical environment matches are not present in the database.

As a further confirmation, the Mann−Whitney test,[29] a nonparametric hypothesis test based on the comparison of two distributions and their median values, was used. This test shows the distributions of residuals represented by mean deviations of 1.85 (direct retrieval) and 1.70 (MLR) ppm to have different medians at a statistical significance of ≫99%, as do the distributions represented by mean deviations of 1.85 (direct retrieval) and 1.69 (PLSR) ppm.

Tests for the Quality of Predicted Chemical Shifts. A weakness of the algorithm used to produce the results discussed above is that if a prediction is performed by use of model building, no tests are performed to judge whether or not the computed model is good enough to be used for the prediction. Many factors can cause a computed model to perform poorly in prediction. For example, the database of 16 959 structures used here is not large enough to ensure that the calibration set of carbons used for model building provides a good match to the environment of the target carbon. This can lead to a model that performs poorly in prediction. Furthermore, any errors in chemical shift assignments in the database can lead to an incorrect dependent variable in the model-building calculation. This can negatively impact the ability of the computed model to produce an accurate predicted chemical shift.

A second issue that affects the quality of the computed models is the accuracy of the environmental encoding algorithm. The current algorithm is very good at finding carbons with closely matching environments. This is reflected in the better results obtained by direct retrieval rather than model building for cases in which an exact or nearly exact match is found in the database. However, as the closeness of the match degrades, the rank ordering of the matches becomes increasingly arbitrary. Thus, the certainty that the top 200 environmental matches have been found also degrades, and the quality of the calibration set used for building the model is reduced. Previously published work on the environmental encoding algorithm discusses some of the specific limitations of the scheme.[16]

The most straightforward procedure for judging whether a model is useful for performing a prediction is to use standard statistical tests for assessing the model quality and to define criteria of acceptability on the basis of these tests. The R^2 value and the standard error of estimate, s (standard error of calibration), were both used to attempt to identify poor models. A further test was based on the Mahalanobis distance[30] of the independent variables describing the target

Table 4. Comparison of Residuals for PLSR Models and Direct Retrieval

	model building	direct retrieval
residual > 10.0 ppm	192	407
residual > 20.0 ppm	22	99
better	1637	1217
better by ≥5.0 ppm	493	165
mean deviation (ppm)	3.76	5.29
median absolute residual (ppm)	2.57	3.30
max absolute residual (ppm)	51.43	55.30

carbon (i.e., the x_i in eq 1) from the distribution formed by the corresponding independent variables of the calibration set. This test judges whether the independent variables describing the target carbon fall within the multivariate distribution of the independent variables used to describe the carbons in the calibration set. The Mahalanobis distance can be converted to a corresponding *F*-statistic and assigned a significance level (probability).[30]

For the best cases from the exploratory experiments for PLSR and MLR, a series of experiments was performed in which model building was not used if a certain R^2 value was not reached (80%, 60%, or none), a certain limit of *s* was exceeded (4, 6, 8, or 20 ppm), or the probability corresponding to the Mahalanobis distance was exceeded ($p = 0.98$ or none). Direct retrieval was always used if the Euclidean distance for the top match was less than 1.0. If neither the direct retrieval nor model-building conditions were met for a given test carbon, the chemical shift for that carbon was simply not predicted.

The best results for both PLSR and MLR were obtained for the most restrictive conditions where R^2 was 80%, *s* was 4.0 ppm, and the probability corresponding to the Mahalanobis distance was 0.98. Of the 4382 (PLSR) or 4372 (MLR) carbons predicted by modeling when no R^2, *s*, or Mahalanobis conditions were applied, 2074 were predicted by PLSR and only 595 were predicted by MLR for the most restrictive conditions. Under these conditions, the mean deviation for PLSR/direct retrieval was 1.40 ppm for the resulting set of 36 196 carbons, and the corresponding mean deviation for MLR/direct retrieval for its set of 34 717 carbons was 1.28 ppm. The apparent success of MLR versus PLSR is at least partially due to the fact that MLR has simply eliminated more of the worst test carbons than has PLSR. If all of the 4382 carbons are removed from the test set, the mean deviation for the remaining 34 122 carbons predicted by direct retrieval is 1.25 ppm.

Good results were also obtained for the case in which the limit of *s* was 4.0 ppm, the Mahalanobis probability was 0.98, and there was no restriction placed on the R^2 value. PLSR/direct retrieval and MLR/direct retrieval gave overall mean deviations of 1.44 and 1.35 ppm, respectively, with the total numbers of carbons predicted using model building being 2855 and 1810, respectively. Thus, for only a small degradation in the mean deviation, significantly more chemical shifts were predicted. The values of the mean deviation obtained if direct retrieval only was used for the test set including the 2855 and 1810 carbons were 1.56 and 1.41 ppm, respectively. When these values are compared to 1.44 and 1.35 ppm, it is seen that PLSR actually performs somewhat better than MLR.

For the set of 2855 carbons noted above, Table 4 provides a more detailed comparison between the results obtained by use of direct retrieval and model building. The first column lists the criterion for the comparison, while the second column gives the results obtained with model building based on PLSR. The third column lists the results that would have been obtained if direct retrieval had been used rather than model building. The first and second rows indicate how many of the 2855 carbons had absolute residuals worse than 10.0 and 20.0 ppm, respectively. The second column in the third row lists the number of carbons for which model building rather than direct retrieval gives a more accurately predicted chemical shift, while the third column gives the number of carbons for which direct retrieval gives better results. The information in the fourth row is the same with the exception that the prediction must be better by at least 5.0 ppm. Rows 5−7 list the mean deviation, median, and maximum absolute residual, respectively, computed across the 2855 carbons.

Table 5. Comparison of Residuals for MLR Models and Direct Retrieval

	model building	direct retrieval
residual > 10.0 ppm	69	180
residual > 20.0 ppm	7	35
better	1038	772
better by ≥5.0 ppm	242	64
mean deviation (ppm)	3.21	4.42
median absolute residual (ppm)	2.35	3.00
max absolute residual (ppm)	37.69	45.00

Table 6. Results of Evaluation of Descriptor Classes

descriptors used	mean deviation (ppm)
topological	1.70
geometrical	1.76
electronic	1.82
topological and geometrical	1.68
topological and electronic	1.70
geometrical and electronic	1.71
topological, geometrical, and electronic	1.69

Table 5 gives results for the analogous MLR experiment in which 1810 carbons were predicted by modeling. The results for both PLSR and MLR clearly show that model building gives superior results to direct retrieval for these cases. The difference between the results for model building and direct retrieval is greater for both the mean deviation and the median absolute residual for PLSR than for MLR. This result provides additional confirmation that PLSR outperforms MLR under the conditions used in this paper.

Study of Descriptors. A series of six experiments was performed to test the utility of the three classes of descriptors (topological, geometrical, and electronic) and their combinations. These experiments were performed with stepwise PLSR for the test set of 38 504 carbons for the same set of conditions which yielded the best results for the full set of descriptors. Table 6 displays the results obtained. The first column describes the descriptors used in each experiment, and the second column records the overall mean deviation for the test set. No R^2, *s*, or Mahalanobis conditions were utilized. The seventh row records the results for the previously cited experiment in which all descriptors were used. It can be seen that all of the results are essentially the same except for the two cases in which geometrical and electronic descriptors are the only class of descriptors used. This shows that with the current set of descriptors, the topological descriptors are most important.

CONCLUSION

The results presented in this paper demonstrate that an automated model-building procedure can be developed and successfully incorporated into a direct retrieval system for ^{13}C NMR spectrum simulation. The addition of the model-building step can significantly improve the ability to estimate chemical shifts when an exact environment match is not present in the spectral database.

There are several areas in which improvements in the algorithm are needed. The key capability which has not yet been fully realized is a reliable estimator of the ability of a model to make accurate predictions. The R^2, *s*, and Mahalanobis distance criteria used to screen models were only partially successful. As indicated in Table 4, despite the use of an *s* cutoff of 4.0 ppm and a strict Mahalanobis distance criterion, 192 of the 2855 carbons met these criteria and still led to absolute residuals greater than 10.0 ppm. Given an improved ability to screen models, the score cutoff for using direct retrieval could be made more stringent and models could be built for a greater percentage of carbons. The model-screening procedure would then be applied to determine if a given model should be used.

A second improvement needed is the refinement of the environmental encoding scheme to be better suited to the retrieval of a subset of carbons when the actual closest chemical environments in the database are not especially close to the environment of the target carbon. Another option is the use of a secondary algorithm to identify and remove poorly matching carbons from the retrieved calibration set.

A more thorough study of the structural descriptors should also be performed. The most useful descriptors should be determined, and descriptors which provide redundant information should be identified and eliminated. Descriptors of types other than the 121 descriptors employed in this work should also be investigated and a larger pool of descriptors used. More sophisticated techniques for screening and subsetting the descriptors could also prove beneficial.

Finally, the availability of a larger database with a greater variety of chemical environments would be useful. The encoding of stereochemical information such as *cis*/*trans* information which is not currently used would also enhance the capabilities of the prediction procedure. The detection of possible incorrectly assigned chemical shifts in the database is also a necessity for better prediction results.

ACKNOWLEDGMENT

The Sadtler Division of Bio-Rad Laboratories, Inc. is acknowledged for providing the spectral and chemical structural data used in this work.

REFERENCES AND NOTES

(1) Bremser, W. HOSE—A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(2) Bremser, W. Expectation Ranges of ^{13}C NMR Chemical Shifts. *Magn. Reson. Chem.* **1985**, *23*, 271−275.
(3) Chen, L.; Robien, W. The CSEARCH-NMR Data Base Approach to Solve Frequent Questions Concerning Substituent Effects on Carbon-13 NMR Chemical Shifts. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 217−223.
(4) Crandell, C. W.; Gray, N. A. B.; Smith, D. H. Structure Evaluation Using Predicted ^{13}C Spectra. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 48−57.
(5) Cheng, H. N.; Kasehagen, L. J. Integrated Approach for ^{13}C Nuclear Magnetic Resonance Shift Prediction, Spectral Simulation and Library Search. *Anal. Chim. Acta* **1994**, *285*, 223−235.
(6) Von der Lieth, C. W.; Seil, J.; Köhler, I.; Opferkuch, H. J. ^{13}C NMR Data Bank Techniques as Analytical Tools. *Magn. Reson. Chem.* **1985**, *23*, 1048−1055.
(7) Chen, L.; Robien, W. OPSI: A Universal Method for Prediction of Carbon-13 NMR Spectra Based on Optimized Additivity Models. *Anal. Chem.* **1993**, *65*, 2282−2287.
(8) Clerc, J. T.; Sommerauer, H. A Minicomputer Program Based on Additivity Rules for the Estimation of ^{13}C-NMR Chemical Shifts. *Anal. Chim. Acta* **1977**, *95*, 33−40.
(9) Jensen, K. L.; Barber, A. S.; Small, G. W. Simulation of Carbon-13 Nuclear Magnetic Resonance Spectra of Polycyclic Aromatic Compounds. *Anal. Chem.* **1991**, *63*, 1082−1090.
(10) 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.
(11) Jurs, P. C.; Ball, J. W.; Anker, L. S.; Friedman, T. L. Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272−278.
(12) Jurs, P. C.; Anker, L. S.; Ball, J. W. In *Computer-Enhanced Analytical Spectroscopy, Vol. 4*; Wilkins, C. L., Ed.; Plenum Press: New York, 1993; Chapter 1.
(13) Stuper, A. J.; Brügger, W. E.; Jurs, P. C. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979; pp 83−90.
(14) Sprague, J. T.; Tai, J. C.; Yuh, Y.; Allinger, N. L. The MMP2 Calculational Method. *J. Comput. Chem.* **1987**, *8*, 581−603.
(15) 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.
(16) Schweitzer, R. C.; Small, G. W. Enhanced Structural Encoding Algorithm for Database Retrievals of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 310−322.
(17) Randic, M. Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(18) Lowe, J. P. *Quantum Chemistry*; Academic Press: New York, 1978.
(19) Greenwood, H. H. *Computational Methods in Quantum Chemistry*; Wiley-Interscience: New York, 1972.
(20) Yates, K. *Hückel Molecular Orbital Theory*; Academic Press: New York, 1978.
(21) Streitwieser, A. *Molecular Orbital Theory of Organic Chemistry*; McGraw-Hill: New York, 1969.
(22) Karplus, M.; Pople, J. A. Theory of Carbon NMR Chemical Shifts in Conjugated Molecules. *J. Chem. Phys.* **1963**, *38*, 2803−2807.
(23) Del Re, G. A Simple MO-LCAO Method for the Calculation of Charge Distributions in Saturated Organic Molecules. *J. Chem. Soc.* **1958**, 4031−4040.
(24) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: Chichester, 1989.
(25) Beebe, K. R.; Kowalski, B. R. An Introduction to Multivariate Calibration and Analysis. *Anal. Chem.* **1987**, *59*, 1007A−1017A.
(26) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(27) Strang, G. *Linear Algebra and Its Applications*; Academic Press: New York, 1976; pp 122−125.
(28) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; Wiley: New York, 1978; Chapter 6.
(29) Hollander, M.; Wolfe, D. A. *Nonparametric Statistical Methods*; Wiley: New York, 1973.
(30) Hand, D. J. *Discrimination and Classification*; Wiley: New York, 1981; pp 122−126.

CI9601731