All of the software described above has been for qualitative determinations. INTG is a simple routine for doing quantitative measurements within these 3D chromatograms. The operator selects the wavelength and the time span of interest and the routine drops perpendicular lines to the baseline and integrates the peak.

CONCLUSIONS

The use of a minicomputer to control data acquisition, reduction, and display in RSS/LC has come a long way from being a difficult and time-consuming task to a relatively simple and routine one. Although additional software ideas exist to further simplify the technique, a working system is at hand, and some of the unlimited applications available can now be studied in detail.[4] Investigations presently underway include separations of vitamins, dyes, and polyaromatic hydrocarbons.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support provided by National Science Foundation Grants GP-35979, CHE76-04321 (H.B.M.), CHE74-02641 (W.R.H), and a U.C. Graduate Council Scholarship and Twitchell Fellowship (M.S.D.).

REFERENCES AND NOTES

(1) M. S. Denton, T. P. DeAngelis, A. M. Yacynych, W. R. Heineman, and T. W. Gilbert, "Oscillating Mirror Rapid Scanning Ultraviolet-Visible Spectrometer as a Detector for Liquid Chromatography", *Anal. Chem.*, **48**, 20–4 (1976).
(2) A. M. Yacynych, Ph.D. Thesis, University of Cincinnati, Cincinnati, Ohio, 1975.
(3) H. B. Mark, Jr., R. M. Wilson, T. L. Miller, T. V. Atkinson, H. Wood, and A. M. Yacynych, "The On-Line Computer in New Problems in Spectroscopy: Applications to Rapid Scanning Spectroelectrochemical Experiments and Time Resolved Phosphorescence Studies" in "Information Chemistry; Computer Assisted Chemical Research Design", S. Fujiwara and H. B. Mark, Jr., Ed., University of Tokyo Press, Tokyo, 1975, pp 3–28.
(4) M. S. Denton and T. W. Gilbert, "Rapid Scanning Liquid Chromatography Detector: Instrumentation, Software and Applications", manuscript in preparation.
(5) A. M. Yacynych and H. B. Mark, Jr., "Automatic Absorbance Calibration Routine for a Computerized UV-Visible Rapid Scanning Spectrometer", *Chem. Instrum.*, in press.
(6) M. S. Denton and T. W. Gilbert, *J. Chromatogr.*, manuscript in preparation.

Interactive Pattern Recognition in the Chemical Analysis Laboratory†

CHARLES L. WILKINS
Department of Chemistry, University of Nebraska—Lincoln, Lincoln, Nebraska 68588

Received June 16, 1977

A proposed interactive organic structure analysis system is described. This system based on the use of an MS-5076 ultra-high-resolution mass spectrometer equipped with both chemical ionization and electron impact sources is proposed to be used in the exploration of a variety of interactive pattern recognition studies aimed toward the development of methods intended to facilitate more data analysis on the laboratory minicomputers which are central elements of the system. An approach whereby a gas chromatograph–infrared interferometer would be linked to the mass spectrometer system and the combined information used as the source of structural inferences is discussed. In particular, the use of factor analysis, simplex pattern recognition, digital learning networks, and search methods are considered. It is suggested that information-theory-based-evaluation methods should be used in selection of the optimum solutions to problems encountered.

During the past few years we have been engaged in research directed toward implementing pattern recognition methodologies in an on-line fashion using laboratory computers. For that reason, we have primarily concentrated on development of methods which would lend themselves easily to adaptation to the laboratory framework. Effectively this has constrained somewhat the variety of pattern recognition techniques which we can realistically consider using. For the present, I will first focus on a proposed chemical analysis system we are in the process of developing and use that as an introduction to discussion of a number of our more recent research efforts. Figure 1 is a block diagram of the proposed analysis system hardware. This diagram contains both elements which are already installed in our laboratory and those which we hope to add in the reasonably near future. As is seen in the upper middle of the diagram, a central element of this structural analysis system is a high-resolution mass spectrometer, the AEI MS-5076 Ultra-High Resolution Mass Spectrometer, which is equipped with both electron impact and chemical ionization sources. In addition, the spectrometer is interfaced both to a gas chromatograph (for rapid mass spectrometric scans of mixture components after separation) as well as a data acquisition and control system which employs a Nova 2/10 computer equipped with 32K 16-bit words of main memory, a cathode ray tube terminal for system control, a 2.5 million word magnetic disk and a rapid electrostatic plotter-printer. This data acquisition system also services a Hitachi RMU-6D medium resolution mass spectrometer which is equipped with a field ionization/field desorption source and an electron impact source as well. Enclosed in the dotted lines in the diagram is the proposed addition to the analysis system. By adding this on-line interferometer-based infrared spectrometer, acquisition of infrared spectra of the same gas chromatographic effluents as are currently routed to the MS-5076 spectrometer would be possible. Because of the need for rapid data reduction, the proposed system would incorporate its own dedicated computer, a rapid plotter, floppy disks for intermediate data storage, and an operator console. As currently visualized, the system would be so constructed that stand-alone operation of the GC–IR system would also be possible. Finally, note the planned link between the mass spectrometer–computer control system and the infrared spectrometer computer. This would be a high-speed digital link which would allow use of the larger system peripherals as well as the transfer of reduced

† Presented at the Joint U.S.–Japan Seminar on "Computer-Assisted Chemical Research Design", Aug 16–20, 1976, Washington, D.C., sponsored by the National Science Foundation and the Japan Society for the Promotion of Science.

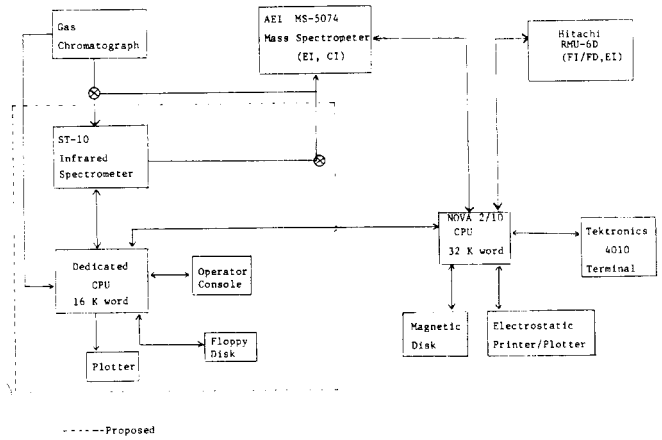

Figure 1. Analysis system hardware.

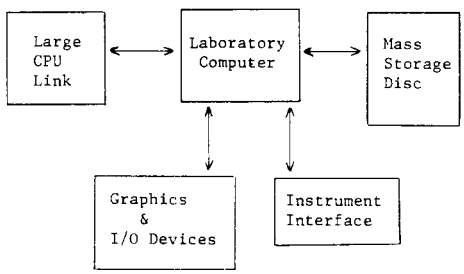

Figure 2. Block diagram of a laboratory pattern recognition system.

data in an effective fashion from the smaller computer.

Most of the research to be described here focuses on the software necessary to most effectively make use of this sophisticated and expensive hardware combination in the laboratory.

The premise of any pattern recognition study directed toward the elaboration of unknown structures from spectral information is that the spectrum is in some definitive way indicative of the structure. Alternately, and the subject of a number of studies by Jurs and Schechter,[1,2] is the converse problem. Namely, knowing the structure, how does one predict the spectrum? Both of these are extremely interesting and challenging problems and neither has been definitively solved. One convenient way of categorizing computer-assisted data interpretation is in terms of human thinking procedures. For example, if one understands the situation entirely, that is to say, is aware of defined analytic solutions to spectrum interpretation, then parametric methods are appropriate. On the other hand, if the analyst believes an unknown spectrum is familiar (i.e., thinks he/she has seen it before), then search methods may be appropriate. In this approach, one allows the computer to match the unknown data, using some suitable algorithm, with a computer-readable base of previously recorded spectra. If satisfactory matches are achieved, then identification of the unknown or its structural elements may be possible. Finally, the analyst may believe that the answer lies within the data but be aware that quantitative analytic solutions are not available to yield structural predictions. In this case, pattern recognition or other empirical techniques are most appropriate. Figure 2 shows a block diagram of the elements of what we visualize to be a general purpose laboratory pattern recognition system. An instrument interface links the laboratory computer with the source of the data to be interpreted. Naturally, a mass storage device is essential if any extensive search procedures are to be possible. A link to a large computer can be extremely useful for similar reasons, although not absolutely essential. The graphics and input/output devices shown in the drawing are particularly important, for they provide a primary means for the analyst to interact with the computer-managed analysis strategies which he may wish to use. As is well known, mass spectra are rich in information and one of the primary problems in mass spectral analysis is to reduce this wealth of information to a usable subsegment. Grotch has quantitatively estimated the information content of mass spectra and has shown that if low-resolution mass spectra with a 200-amu range are considered and simply coded in a binary fashion ("1" for the presence of a peak, "0" for the absence of a peak) that the maximum possible information would be 2^{200} bits (or 2×10^{60} bits).[3] This enormous potential information, of course, is not realized since, in addition to the fact that only certain fragmentation processes are allowable and therefore not all mass positions enjoy equal probability of peaks appearing, atomic compositions also make certain m/e values improbable. It is still clear that the amount of information contained in a mass spectrum is enormous. One of the purposes of our research is to effectively constrain the data interpretation problem as much as possible, using much the same philosophy as Sasaki, for example, has successfully demonstrated with some of his search-based methods.[4,5] In particular, we intend to use gas–liquid chromatography as a means of obtaining preliminary information about both the number and kinds of materials present in unknown mixtures. We will use factor analysis based techniques for determining the number of components in unknown gas chromatograph effluents[6] and retention indices will provide preliminary information on possible functional group composition.[7-9] The infrared spectra of unknowns will provide the first clue about the functional groups present in the unknown species being analyzed, and mass spectrometry will provide fragmentation patterns, elemental analysis, and confirming evidence about functional groups. In order to do this, we expect to use a variety of different algorithms. We will not restrict ourselves to the use of pattern recognition algorithms, but since space does not permit a more lengthy discussion, this paper will concentrate on the pattern recognition algorithms we expect to use in the system. First, the linear learning machine method, which has been the subject of numerous investigations over the past several years,[10-13] will be used because it provides a type of readily implemented discriminant function suitable for rapid application using a laboratory computer. A new technique, which seeks improved linear discriminant functions by a well-known optimization procedure, simplex pattern recognition,[14] will serve as a second source of these functions. Finally, digital learning networks,[15-18] which are multicategory classifiers (in contradistinction to the previous two types of classifiers which are essentially binary classifiers), are under active examination to determine their suitability for use in the analysis system. Although many may be familiar with some or all of these techniques, the fundamentals will be very briefly reviewed here for the sake of completeness.

LINEAR LEARNING MACHINE

The data that are used in pattern recognition are represented as pattern vectors of the form $X = (X_1, X_2, \ldots, X_d)$. The dimensionality, d, of the vector tells the number of observations or features that are used to characterize the pattern. The quantities X_i are the numerical values of each observation i. For binary classifiers, each pattern is classified into one of two categories. In this case, the pattern recognition problem is to determine the relationship between the data and the categories of the data. Figure 3 is a plot of melting point vs. boiling point for several ketones and carboxylic acids. This diagram, which was first suggested by Isenhour and Jurs,[19] provides a simple low-dimensional example to illustrate the principle of linear discriminant functions. As shown in the diagram, it is possible to draw a line separating the acid category (A) from the ketone category (K). This discriminant

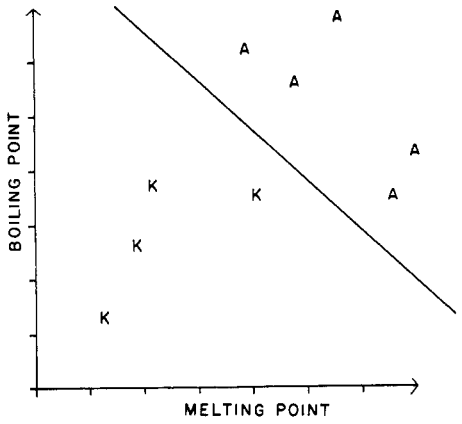

Figure 3. A plot of melting point vs. boiling point for ketones and acids.

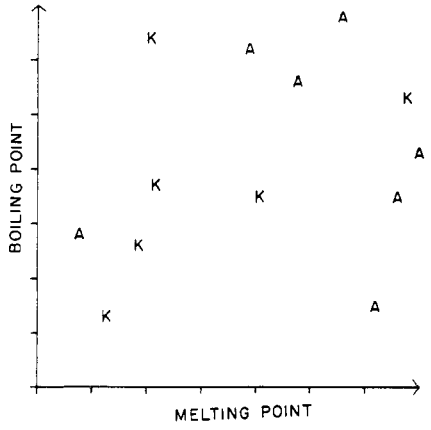

Figure 4. A set of melting point vs. boiling point data which are inseparable by a single linear discriminant.

function geometrically represents a hyperplane, of the same dimensionality as the data, which divides the data into the two desired categories. Points on one side of the hyperplane always belong to category one and points on the other side to category two. If such a hyperplane exists, the data are said to be linearly separable and can be separated by a linear discriminant function of the form

$$S = \sum_{i=1}^{d+1} w_i \cdot X_i \qquad (1)$$

where X_i is the ith component of X and w_i is the weight assigned to that component. A $(d + 1)$ component is added $(X_{d+1} \equiv 1)$ so that the category is determined by the sign of S. For example,

$S > 0$ implies category 1

$S < 0$ implies category 2

The vector formed by the set of weights $(w_1, w_2, \ldots, w_{d+1})$ is the weight vector.

However, one problem may arise, as shown in Figure 4, when additional points are considered. Here, with this example, it is clear that the data are no longer linearly separable. That is to say, a single straight line which will separate the two categories cannot be drawn. Figure 5 shows a number of possible lines which could be tried in an attempt to categorize the largest possible number of members of the two categories. In this situation, the desired outcome of a learning machine computation of a discriminant function would be to obtain that function which correctly categorizes the largest number of the training set.

The algorithm used to calculate the weight vector parallels that first described by Rosenblat in 1960.[20] The method

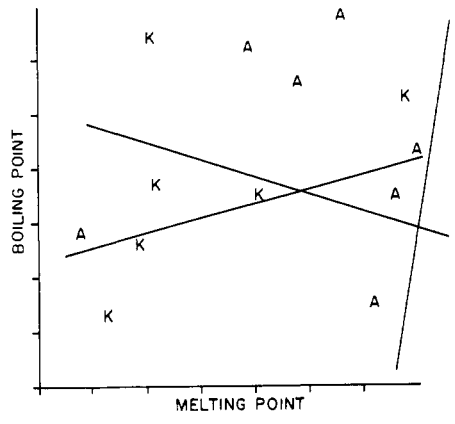

Figure 5. Trial linear discriminants for inseparable data.

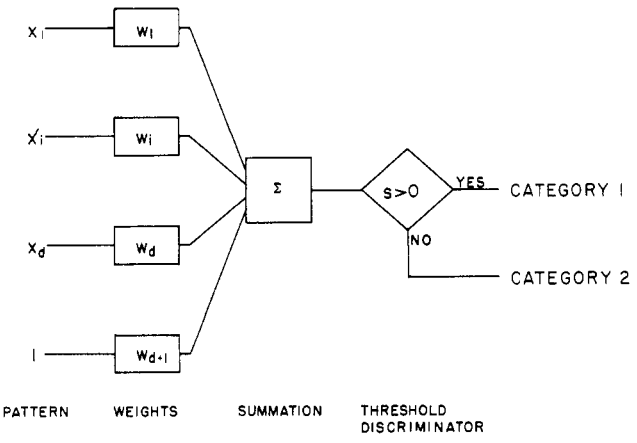

Figure 6. A threshold logic unit.

employes an error-correction feedback procedure (i.e., negative feedback for an incorrect response) and has appeared frequently in the chemical literature. This algorithm must converge to a solution if the data are linearly separable, although the *rate* of convergence cannot be predicted. If the data are linearly inseparable, no weight vector solution exists and the weight vector fluctuates greatly in the vector space formed by the weights. Thus, for inseparable data such as those of the Figure 4, the normal procedure will never allow convergence, and the weight vector has much less utility as a classifier of unknowns. Weight vectors are generally calculated using as large and representative a training set as possible in an attempt to ensure that they will perform adequately with unknown data. Because of this, large computers are employed by us and others to obtain weight vectors which may then be used subsequently in laboratory computer systems. Because of the sometimes lengthy computation and large memory requirements for efficient calculation, it has not generally been found practical to use laboratory computers in this phase of work. Let us now turn to how one would employ this procedure in chemical pattern recognition analysis.

As an example, consider how a mass spectrum could be depicted. A typical presentation of such a spectrum would be a graph of m/e values vs. peak intensities. As is obvious, one way to encode the spectrum would be to use a series of 0's and 1's equal in number to the number of resolution elements in the spectrum, with a 1 being used to indicate the presence of one or more peaks in an interval and a 0 to indicate the absence of a peak. If one uses a set of such encoded spectra to train a linear discriminant to answer some particular structural question (i.e., is a halogen present or absent in the structure of the compound which generated the spectrum?), then the linear discriminant function (weight vector) developed can be used to categorize unknowns. As is shown in Figure

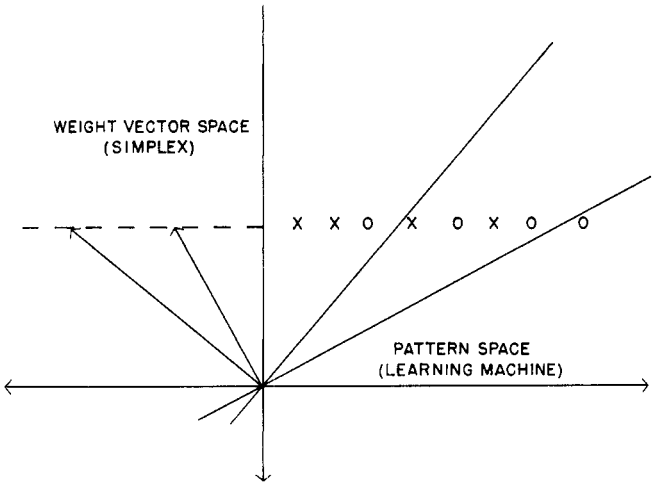

Figure 7. The relationship of weight vector and pattern space for a one-dimensional binary categorization.

6, the encoded spectrum, referred to in the diagram as the pattern, is simply represented as a vector which is multiplied by the weight vector in order to obtain a product (called "summation" in the figure), and a threshold discriminator examines this sum, compares it with 0, and assigns the unknown to category 1 if the sum is greater than 0 and category 2 if it is less than 0. As is obvious, the use of such a discriminant procedure requires minimal memory storage and is quite rapid. If one could obtain the best possible discriminants and their levels of performance were sufficiently high to allow a measure of reliance on their predictions, they should be of significant value in the structural analysis system.

However, as mentioned before, linear inseparability is likely with real spectral data, if representative numbers of spectra are considered. Accordingly, there is a need for a technique which will search weight vector space in an attempt to find the maximum in the recognition response curve. Simplex pattern recognition is one such method.

SIMPLEX PATTERN RECOGNITION

This optimization technique was first applied to the optimization of linear discriminant functions for chemical application by us in collaboration with Isenhour and his students.[14] The method, which will be described briefly, is one which has been applied extensively in experiment optimization problems. As we employ it in pattern recognition, the method searches the weight vector space in order to find an optimum weight vector. Figure 7 illustrates this procedure with a one-dimensional pattern separation problem. Here the X's and 0's are to be separated by a linear discriminant function. It is obvious that these are inseparable by a single such function and the problem is to find that weight vector which will make the fewest mistakes. Two such weight vectors are depicted in the diagram. It is clear that the one to the left is capable of correctly categorizing half of the x's and one-fourth of the 0's. On the other hand, the second categorizer would correctly categorize all of the X's at the expense of miscategorizing three-fourths of the 0's. In any case, the point is that the simplex method will work in that region of the space referred to as the weight vector space, rather than the pattern space used in the linear learning machine method. The sequential simplex method we have used is that originally proposed by Spendley, Hext, and Himsworth[21] and later modified by Nelder and Mead.[22] This approach to optimization problems is geometrically appealing and has been applied to date to two types of chemical pattern recognition analysis. It has been applied in our laboratory to both carbon-13 NMR spectral interpretation[23] and mass spectrometric analysis.[24] As is shown

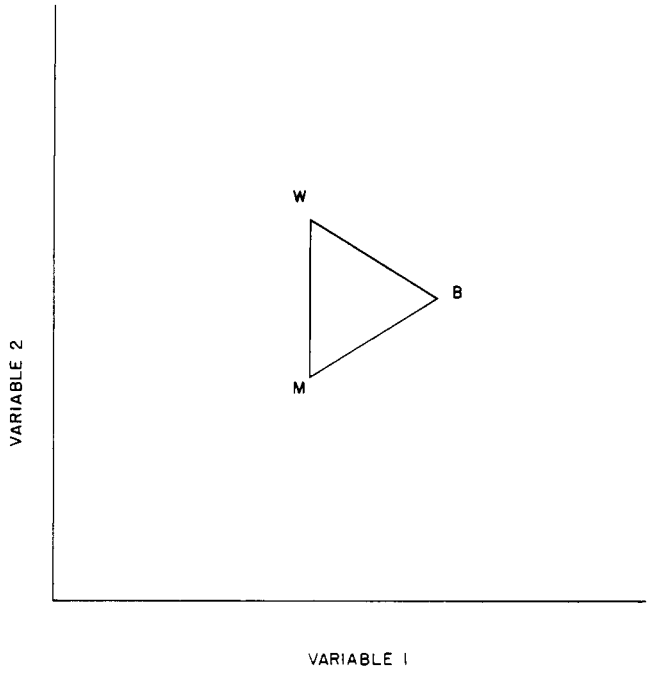

Figure 8. A two-dimensional Simplex.

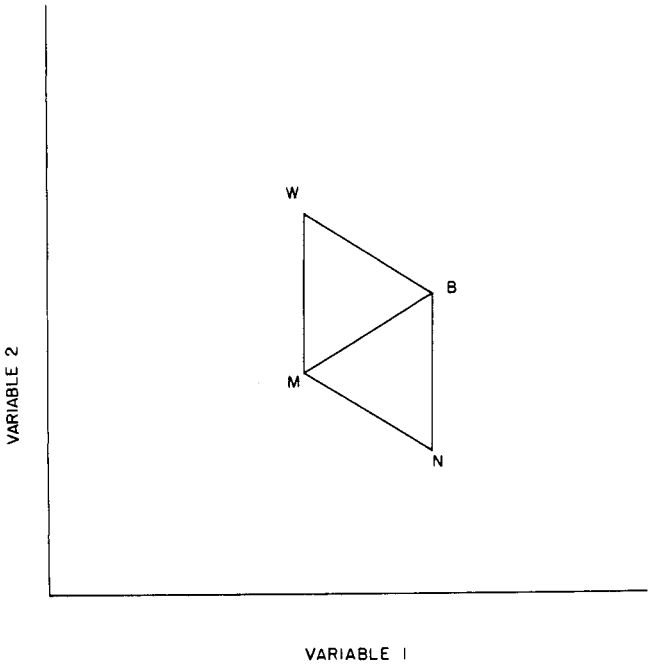

Figure 9. Movement of the two-dimensional Simplex by reflection.

in Figure 8, a simplex is a geometric figure used in the optimization procedure. If the optimization is to be done over d variables, the simplex will contain $d + 1$ vertices in d-dimensional variable space. For example, a two-dimensional simplex such as that in the diagram is a triangle and the three-dimensional simplex is a tetrahedron.

A response function is evaluated for each of the vertices. In the diagram, the vertex with the best response is labeled B, that with the worst labeled W, and that of intermediate response labeled M. Then, the simplex is moved along the response surface in weight vector space to find an optimum. This optimum is approached by movement away from the least desirable response, as illustrated in Figure 9. In its original form, the simplex moved by a direct reflection away from the worst response across the other d vertices. Modifications by Nelder and Mead[22] allow the simplex to follow more closely the contours of the response surface. These modifications are

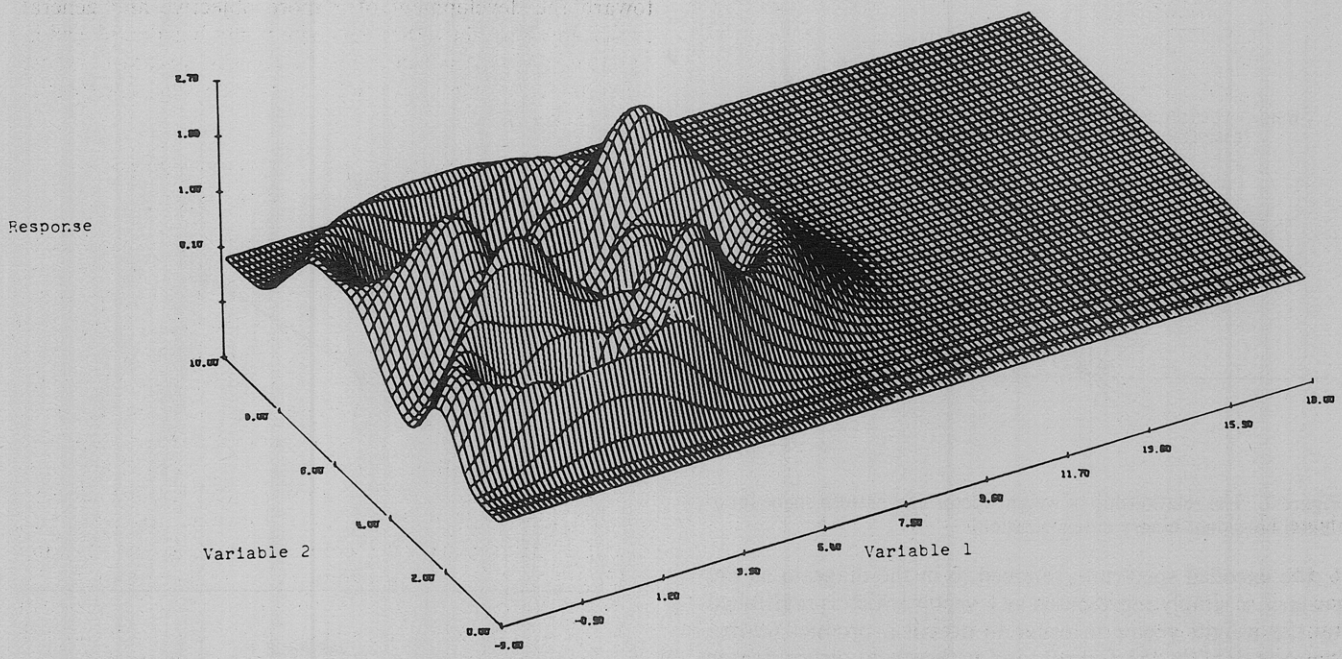

Figure 10. A hypothetical response surface for a two-variable experiment.

minor and will not be further discussed here. In order to apply the simplex optimization method, it is usually assumed the response surface has a unique extremum and is continuous. Figure 10 shows an example of a typical response surface. As can be seen, there are a number of maxima in the sample surface shown, and it would be possible to locate a local rather than a global extremum. The latter, of course, is the desired outcome. By the specific nature of the response in the pattern recognition problem, it is discontinuous; that is, only integral numbers of patterns may be classified. Therefore, it is possible for the simplex to become stranded on a plateau (i.e., an equi-recognition region). To force the response surface to be continuous, a second optimization criterion is added. In the studies discussed here, the perceptron criterion function has been used as a secondary criterion. This forces the response to have the form of a continuous variable. The best weight vector will then be that which has the minimum perceptron function value and the maximum recognition. To summarize the steps in simplex pattern recognition: (1) $n + 1$ weight vectors are selected (n is the dimension of the vector space); (2) all patterns are classified with each vector; (3) the vector with the worst response is reflected; (4) the simplex is expanded or contracted when appropriate. All patterns are then classified with the new vector and the reflection procedures repeated. This iterative approach continues until no further improvement occurs. At that point the weight vector with the best performance is assumed to be in hand and is applied as shown in Figure 6. That is to say, the encoded pattern (the mass spectrum or infrared spectrum) is simply multiplied by the weight vector and a threshold discriminator is then employed to assign the unknown to one of two possible categories. It should be obvious at this point that *application* of simplex-derived weight vectors is no different from application of those obtained by the simpler linear learning machine method.

ADAPTIVE DIGITAL LEARNING NETWORKS

Recently, the machine recognition of chemical classes for mass spectra using a multicategory classifier has been described by Stonham and co-workers.[15,16] This approach, the adaptive digital learning network, stems from the earlier work of Bledsoe and Browning[17] on n-tuple sampling for pattern recognition. Its application requires use of only class members for training and binary patterns as the representation of spectra. The early reports indicated this approach could be significantly superior to pattern recognizers previously applied to chemical data. For that reason, we have examined the digital learning network approach.

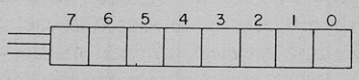

Figure 11. An 8-bit digital learning element.

The basic unit of an n-tuple learning machine is the digital learning element, which may be thought of as a storage register, as shown in the Figure 11. This storage register is associated with a *randomly chosen* group of n pattern elements from the pattern being presented to the learning machine for training or for recognition/prediction. This group is an n-tuple subpattern. The number of storage locations in the storage register is 2^n where n is the number of pattern elements in the n-tuple. Thus, a 3-tuple sampling requires a $2^3 = 8$ bit register, as in the diagram. For this 3-tuple sample, there are 8 possible configurations, (000), (001), (010), . . ., (111). The configuration or bit pattern of the n-tuple subpattern is used to address one of the 2^n locations in the digital learning element restorage register. In the training stage, a "1" is written into the bit position addressed by the n-tuple subpattern. In the recognition/prediction stage, the bit is read rather than written. A "digital learning network" consists of a group of digital learning elements. The number of elements used in the network depends on the n-tuple value (n), on the number of pattern features being examined, and on the sampling redundancy desired. For example, if 90 pattern features are to be subjected to a 3-tuple sampling with no feature being sampled twice, then 30 8-bit learning elements will be required to make a learning network. One learning network is prepared for each category of patterns to be classified.

To train a digital learning network for a particular category, a pattern belonging to that category is presented to the network, which has been initialized to contain all "0"'s. Each learning element in the learning network is associated with its own n-tuple subpattern in the pattern being studied, and the bit pattern in that n-tuple subpattern addresses just one of the 2^n bit positions in the learning element. In each location

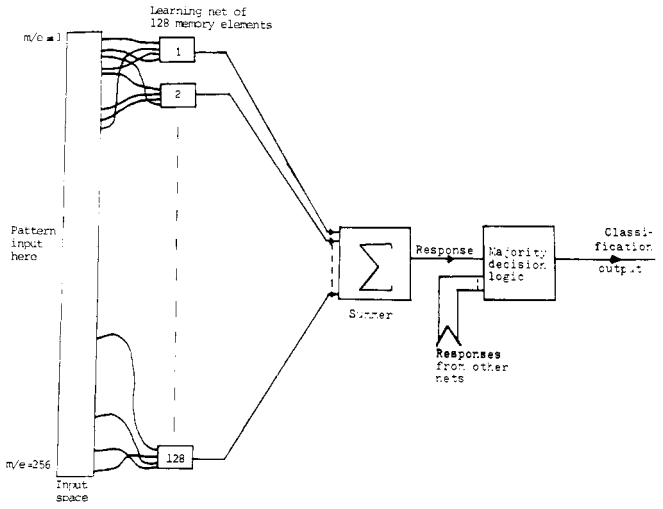

Figure 12. Block diagram of a digital learning network.

so addressed, a "1" is now written. Now, there will be just one "1" in each learning element in the network. Further training proceeds by presenting to the network more patterns from the given category. When recognition of a pattern is desired, the pattern is presented to the network just as for training; the *n*-tuple subpatterns are used to address locations in the learning elements comprising the learning network. However, instead of *writing* into the memory array, the contents of the locations being addressed are now *read* and those values summed. Any *n*-tuple subpattern which was present in a training pattern will thus address a location which had been written with a "1" during training; *n*-tuple subpatterns which were not present in any training pattern will address locations which are unchanged from their initial "0" value. Figure 12 illustrates the logical connection of the elements within a classifier. As is seen in this drawing, an unknown pattern is submitted sequentially to each of several digital learning networks and a response is obtained for each one. Comparison logic is then applied to determine which of the networks obtain the highest response and what the correct classification should be. The network with which the pattern achieves the highest score gives the predicted class identity of the pattern.

ON-LINE STRUCTURE SYSTEM REQUIREMENTS AND PRELIMINARY RESULTS

Primary in an on-line structure system is the requirement that rapid and efficient algorithms be developed in order that analyst may obtain results which are both helpful and not inordinately consuming of either the laboratory computer's, facilities, or the analyst's time. A variety of strategies such as linear discriminant function use, mapping procedures, and search techniques ought to be available to the analyst. Spectrometrists must be provided with rapid and effective means of interaction with the analysis system and complementary information should be utilized wherever possible to constrain the analysis problem. One of the difficulties in examining alternate strategies has been the lack of suitable objective procedures for pattern classifier evaluation. This is particularly true since studies reported in the literature have often been carried out with diverse data sets and in various laboratories. A primary point to consider is the need to maintain the distinction between recognition and prediction. Ideally, whatever procedure is used to evaluate alternate classifiers for use in a structure analysis system should be independent of the classification method and the test data set composition. Therefore, one of the main aspects of our research in the past several months has been effort directed toward the development of a more objective and general evaluation procedure than has commonly been employed in the past. This procedure will be briefly discussed before turning to the results of the preliminary studies in mass spectral interpretation.

NEW APPROACH TO EVALUATION OF ALTERNATE CLASSIFIERS

Because the primary goal of the proposed research is to develop working on-line pattern recognition systems for chemical data analysis, it is essential that we adopt some objective means of comparing classifiers developed in different ways or having different properties. Only in this way will it be possible to judge anticipated levels of performance within the laboratory framework. Both previous studies by ourselves and others, as well as those discussed here, have employed classifiers developed in different ways, including linear discriminant functions, minimum distance measures, and adaptive learning networks. As Uhr pointed out some time ago, ". . . Indeed, there is virtually no comparison evidence for any pairs of programs. . .".[25] He attributed this, in large part, to the lack of suitable standard format test data. Another problem is the lack of convenient comparison algorithms. In the present case, since it is desired that the comparison test set reflect as well as possible the "universe" of unknown patterns to which classifiers will be applied in actual use, the largest possible set should be employed. Equally important, for comparison purposes, is that a *common* test be used with all categorizers being evaluated. It is, of course, understood that application of comparative results to unknown data will be successful to the degree that the data resemble the test set.

Recent theoretical work by Rotter and Varmuza[26] has suggested a promising method of obtaining suitable measures of classifier performance with diverse classifiers. Their information-theoretic analysis appears to offer the basis of the sought-for comparison method. We propose to utilize a practical evaluation procedure we have recently derived from their theoretical results. Because our derivation of the procedure has not been published at this writing, a summary of its basis and the derivation itself are included here.[28] We believe use of this approach may provide the necessary realistic and objective comparison necessary to interpret comparative studies of adaptive learning networks, simplex, and linear learning machine categorizers in mass spectral and NMR interpretive systems.

Rotter and Varmuza suggest that for binary classifiers either predictive ability for both classes or the information gain of the classifiers are suitable objective criteria for comparison. Information gain, $I(A,B)$, is defined as the difference in entropy (H) (basically uncertainty in class membership) before and after application of the classifier. Equations 2–7 summarize the pertinent relationships.

The entropy before applying the classifier, $H(A)$, is computed using the a priori probabilities, $p(1)$ and $p(2)$, of a member of the set belonging to either class 1 or class 2 (eq 2). Since base 2 logarithms are used, the units of both entropy and information are bits. Similarly, the entropy after application of the classifier, $H(A|B)$, can be calculated from the conditional entropies for categorizations of class 1 (j), or class 2 (n), and the a priori probabilities for class 1 or 2 categorizations by the classifier, using eq 3–5.

$$H(A) = -p(1)\log_2 p(1) - p(2)\log_2 p(2) \quad (2)$$

$$H(A|j) = -p(1|j)\log_2 p(1|j) - p(2|j)\log_2 (2|j) \quad (3)$$

$$H(A|n) = -p(1|n)\log_2 p(1|n) - p(2|n)\log_2 (2|n) \quad (4)$$

$$H(A|B) = p(j)H(A|j) + p(n)H(A|n) \quad (5)$$

Information gain, $I(A,B)$, of the classifier is then simply the difference in the entropies calculated in eq 2 and 5.

$$I(A,B) = H(A) - H(A|B) \tag{6}$$

An alternate formulation of the information gain can be derived:

$$I(A,B) = \sum_{i=1,2} \sum_{k=j,n} p(i,k) \log_2 \frac{p(i,k)}{p(i)p(k)} \tag{7}$$

This equation is particularly useful for the experimental application of the measure, since all of the quantities necessary for its solution can be computed from four easily tabulated experimental quantities. These are:

N_i = the number of members of class i

N_i^{pred} = the number of patterns predicted to be in class i

N_i^{corr} = the number of patterns correctly predicted to be members of class i

N_{total} = the total number of patterns in the test set

The required relations are summarized in eq 8–15:

$$p(1) = \frac{N_1}{N_{total}}; \; p(2) = 1 - p(1) \tag{8,9}$$

$$p(j) = \frac{N_i^{pred}}{N_{total}}; \; p(n) = 1 - p(j) \tag{10,11}$$

$$p(1,j) = \frac{N_i^{corr}}{N_{total}}; \; p(1,n) = \frac{N_i - N_i^{corr}}{N_{total}} \tag{12,13}$$

$$p(2,j) = \frac{N_i^{pred} - N_i^{corr}}{N_{total}};$$

$$p(2,n) = \frac{N_{total} - N_i^{pred} - N_i - N_i^{corr}}{N_{total}} \tag{14,15}$$

Because the amount of possible information gain is a function of the test set distribution (eq 2), it is obvious that valid comparisons between classifiers require a common test set of identical distributions of class membership if this measure is to be used. For qualitative purposes, a dimensionless quantity such as the ratio or information gain to a priori entropy, eq 16, might be useful.

$$\frac{I(A,B)}{H(A)} = \frac{\text{information gain}}{\text{original uncertainty}} = M \tag{16}$$

Although the information gain parameter (M) will be of clear value in helping decide which of several alternate classification procedures to use to answer particular questions or to aid in evaluation of the merits of alternate preprocessing or feature selection procedures, of more interest to chemists using the classifiers are the a posteriori prediction capabilities. That is, given that the categorizer predicts a particular class membership for an unknown, what is the expectation that the prediction is correct? This is somewhat different from the more commonly reported statistic, the class conditional probability (i.e., given that a pattern belongs to a particular class, what is the expectation that it will be correctly categorized?). We propose to provide a posteriori probabilities derived from a large test set as part of the information which will be supplied to the scientist with predictions from the on-line systems. Use of such probabilities must be tempered with judicious care, since they can be misleading. A hypothetical example is a collection of 1000 members with 999 in one class and 1 in the second. If both classes were correctly predicted by the classifier, the a posteriori probability, $p(1|j)$, would be 100%. It would be clearly erroneous to assign this probability as a figure of merit to all future predictions of that class. Obviously careful attention must be paid to test set statistics and composition. Because of this sensitivity to test set composition, such probabilities should be most useful in carefully constrained applications.

Table I. 1252 Compound Data Set

Category Number	Compound Class	Number of Spectra
1	Arenes	249
2	Aldehydes and ketones	96
3	Ethers	103
4	Aliphatic alcohols	185
5	Phenols	84
6	Carboxylic acids	51
7	Thiols	135
8	Esters	125
9	Amines	131
10	Amides	56
11	Nitriles	37

Table II. Comparison of Typical Simplex, LLM, and DLN Mass Spectrum Classifiers

Functional Group	Best Linear Discriminant[A]			Best DLN[A]		
	P(J|1)	P(N|2)	M	P(J,1)	P(N,2)	M
Phenyl	92	95	63(20 sim)	98	93	69
Ether	77	85	21(60 LLM)	32	99	20
Nitrile	88	97	49(60 LLM)	100	99	87
Acid	71	9[illegible]	25(60 sim)	16	99	7
Amines	95	94	57(60 LLM)	61	95	29

[A] All numbers are percentages.

In terms of the previously mentioned experimental quantities, both class conditional and a posteriori probabilities are easily obtained. Equations 17 and 18 show these relationships.

$$p(j|1) = \frac{N_i^{corr}}{N_i} \quad \text{class conditional probability} \tag{17}$$

$$p(1|j) = \frac{N_i^{corr}}{N_i^{pred}} \quad \text{a posteriori probability} \tag{18}$$

COMPARISON OF CLASSIFIERS BASED ON LINEAR LEARNING MACHINE, SIMPLEX, AND DIGITAL LEARNING NETWORK METHODS

As may be seen from Table I, for the present studies we used a large diverse mass spectral data set which contained spectra of compounds representing 11 different categories of structural features. This 1252 compound data set formed the basis for the studies whose results are reported in Tables II and III.[24] Here, when typical simplex, linear learning machine, and digital learning network classifiers are compared for a number of a functional group questions (all the result of application of the classifiers to more than 1000 unknown spectra), a number of conclusions clearly emerge. First, for the functional groups considered, the figure of merit must be considered in conjunction with the class conditional probabilities. For example, if the ether question of Table II is considered, it can be seen that a 60-feature linear learning machine discriminant had a merit figure of 21% and the best digital learning network had a merit figure of 20%. However, in the later case, this was a result of a very high class conditional probability for the absence of the ether function, whereas in the linear learning machine case the merit figure reflected a rather low but

Table III. Comparison of Linear Mass Spectral Discriminants with STIRS[A] Results

FUNCTIONAL GROUP	BEST LINEAR DISCRIMINANT[B]			STIRS[B]		
	P(J 1)	P(N 2)	M	P(J 1)	P(N 2)	M
ALDEHYDE AND KETONE	87	96	50(60 LLM)	30	100	18
ALCOHOL	82	89	34(60 LLM)	42	97	20
PHENYL	92	95	63(20 SIM)	75	95	44
THIOL- THIOETHER	93	95	55(20 SIM)	35	99	22
ETHER	77	85	21(60 LLM)	50	97	26

[A] H.E. DAYRINGER, G.M. PESYNA, R. VENKATARAGHAVAN, AND F.W. MCLAFFERTY, ORG. MASS SPECTR., 11, 529 (1976).

[B] ALL NUMBERS ARE PERCENTAGES

approximately similar performance as reflected by the class conditional probabilities. It is also seen, from further examination of this table, that there are cases where simplex discriminant is best, others where the linear learning machine weight vector performs best, and still others where the digital learning network is superior.

These results suggest that none of these classification techniques is superior and that all of them ought to be available in the structure analysis system. Table III presents the results of comparison of the best linear discriminants obtained in our study with results from the STIRS study carried out by McLafferty and co-workers.[27] It should be noted in considering these results that there are a few differences which make them not strictly comparable, although the use of merit figures and class conditional probabilities makes it possible to reasonably well compare them. Our study used 1052 monofunctional unknowns in obtaining the figures presented in the table while the STIRS study employed 373 unknowns which were not restricted to monofunctional compounds. It is seen that the STIRS results are generally unbalanced for the functional groups we have listed here. That is to say, high class-conditional probabilities result for the nonmember predictions, but significantly worse results are obtained for spectra of compounds containing the groups in question. In particular, the balanced performance we deem desirable is lacking in the STIRS results. This is generally reflected in the merit figures although again in the case of ether, we see an example where a higher merit figure is obtained as the result of one high class conditional probability combined with a much lower one. This reemphasizes the need for caution in comparing classifiers based upon only one evaluation criterion.

CONCLUSIONS

As a result of our studies to date, it is apparent to us that a variety of discriminant functions should be included in the structure analysis system. It is equally clear that no single pattern recognition approach will be sufficient but that a number of different approaches are probably to be desired. Although space does not permit a discussion of search approaches, it is also clear that capabilities of this sort should definitely be available in the system.

It is further obvious that structure determination using mass spectra alone is unlikely to reach a level of 100% accuracy, and therefore, the strategy of employing complementary *information*, much as we intend to employ complementary data analysis methods, will be particularly vital.

Finally, the spectrometrist will continue to play a vital role in spectrum interpretation since it will be necessary for the analyst to choose from among a variety of analysis methods and to decide how to best proceed in moving from presentation of the original analysis problem toward the final structural solution. We are confident, however, that the ultimate goal of a semiautomated structure analysis system deriving its data from a number of types of physical measurements is a realistic and possible goal and that it will be achieved within the foreseeable future.

ACKNOWLEDGMENT

The financial support of the National Science Foundation under Grant MPS 74-01249 is gratefully acknowledged.

REFERENCES AND NOTES

(1) J. Schechter and P. C. Jurs, *Appl. Spectrosc.*, **27**, 30 (1973).
(2) J. Schechter and P. C. Jurs, *Appl. Spectrosc.*, **27**, 225 (1973).
(3) S. L. Grotch, *Anal. Chem.*, **42**, 1214 (1970).
(4) S. Sasaki, *Pittsburgh Conf. Abstr.*, No. 396 (1976).
(5) Y. Kudo and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, **16**, 43 (1976).
(6) G. L. Ritter, S. R. Lowry, T. L. Isenhour, and C. L. Wilkins, *Anal. Chem.*, **48**, 591 (1976).
(7) H. Nau and K. Biemann, *Anal. Chem.*, **46**, 426 (1974).
(8) E. Kovats, *Helv. Chim. Acta*, **41**, 1915 (1968).
(9) L. S. Ettre, *Chromatographia*, **7**, 261 (1974), and preceding papers cited therein.
(10) T. L. Isenhour, B. R. Kowalski, and P. C. Jurs, *CRC Crit. Rev. Anal. Chem.*, pp 1–44 (July 1974).
(11) B. R. Kowalski, *Comput. Chem. Biochem. Res.*, **2**, 1–76 (1974).
(12) P. C. Jurs and T. L. Isenhour, "Chemical Applications of Pattern Recognition", Wiley, New York, N.Y., 1975.
(13) P. C. Jurs, "Proceedings of the Workshop on Chemical Applications of Pattern Recognition", Department of Chemistry, The Pennsylvania State University, University Park, Pa., 1975.
(14) G. L. Ritter, S. R. Lowry, C. L. Wilkins, and T. L. Isenhour, *Anal. Chem.*, **47**, 1951 (1975).
(15) T. J. Stonham and M. A. Shaw, *Pattern Recognition*, **7**, 235 (1975).
(16) T. J. Stonham, I. Aleksander, M. Camp, W. T. Pike, and M. A. Shaw, *Anal. Chem.*, **47**, 1817 (1975).
(17) W. W. Bledsoe and I. Browning, "Pattern Recognition and Reading by Machine", reprinted in "Pattern Recognition", L. Uhr, Ed., Wiley, New York, N.Y., 1966, pp 301–316.
(18) L. J. Soltzberg, C. L. Wilkins, S. L. Kaberline, T. F. Lam, and T. R. Brunner, *J. Am. Chem. Soc.*, **98**, 7144 (1976).
(19) T. L. Isenhour and P. C. Jurs, *Anal. Chem.*, **43**, (10), 20A (1971).
(20) N. J. Nillson, "Learning Machines", McGraw-Hill, New York, N.Y., 1965.
(21) W. Spendley, G. R. Hext, and F. R. Himsworth, *Technometrics*, **4**, 441 (1962).
(22) J. A. Nelder and R. Mead, *Comput. J.*, **7**, 308 (1965).
(23) T. R. Brunner, C. L. Wilkins, T. F. Lam, L. J. Soltzberg, and S. L. Kaberline, *Anal. Chem.*, **48**, 1146 (1976).
(24) T. F. Lam, C. L. Wilkins, T. R. Brunner, L. J. Soltzberg, and S. L. Kaberline, *Anal. Chem.*, **48**, 1768 (1976).
(25) L. Uhr, "Pattern Recognition", John Wiley and Sons, Inc., New York, N.Y. (1966), p 377; see also, his more recent comments on the general problem of evaluation in L. Uhr, "Pattern Recognition, Learning, and Thought", Prentice-Hall, Inc., Englewood Cliffs, N.J. (1973) pp 26–28.
(26) H. Rotter and K. Varmuza, *Org. Mass Spectrom.*, **10**, 874 (1975).
(27) H. E. Dayringer, G. M. Pesyna, R. Venkataraghavan, and F. W. McLafferty, *Org. Mass Spectrom.*, **11**, 549 (1976).
(28) NOTE ADDED IN PROOF. This has now been published in more detail: L. J. Soltzberg, C. L. Wilkins, S. L. Kaberline, T. F. Lam, and T. R. Brunner, *J. Am. Chem. Soc.*, **98**, 7139 (1977).