Joint Neural Network Interpretation of Infrared and Mass Spectra

Christoph Klawun[‡] and Charles L. Wilkins*

Department of Chemistry, University of California, Riverside, California 92521-0403

Received August 29, 1995[⊗]

Combining gas phase infrared (IR) spectra with mass spectral (MS) data, a neural network has been developed to predict 26 different molecular substructures from multispectral information. The back-propagation procedure has been used for training, including its previously published modification, the flashcard algorithm. Present functional groups have been detected correctly in 86.4% of all cases, compared with 88.4% using only IR and 78.2% using only MS data for training and prediction. For only 8 out of the 26 functionalities does the joint utilization of infrared and mass spectra yield better prediction results, with the greatest improvement being for halogen bond predictions. The prediction of functional group absence results in accuracy of about 95.5% for both IR and IR/MS networks but only 87.1% for a stand alone MS network. Insights have been gained into the suitability of both data sets for neural network training by presenting just IR or MS data to a jointly trained neural network, revealing the amount of information the network utilizes from either spectroscopic technique. In addition, an algorithm which produces balanced training and test sets for multi-output neural networks has been devised.

1. INTRODUCTION

Ever since Rumelhart and co-workers introduced the back-propagation algorithm in 1986,[1,2] it has been widely used for neural network spectral interpretation in analytical chemistry. Notably, the analysis of infrared spectra attracted a number of research activities, starting with the first papers by Munk and his students.[3,4] With the broad scope of Munk's work, a number of training and prediction problems surfaced and were examined subsequently by other researchers. The impact of different network architectures, spectral resolution, and overtraining effects on prediction quality was examined by Weigel and Herges,[5] and attempts were made to reconstruct IR spectra for selected functional groups from the network weights.[6,7] Neural networks were also implemented as a storage and retrieval system for 1129 IR spectra[8,9] and successfully served as a fast library search prefilter for 609 matrix isolation IR spectra.[10] Fessenden and Györgyi[11] studied the effect of data representations in the cm^{-1} or μm domains with absorbance or transmission intensities and varying number of input data points. They found that using percent transmission in the μm domain with 462 data points provided the best training conditions. Improved prediction quality resulted from assigning individual networks to recognize specific functional groups instead of training a single large network for this task.[12,13]

For mass spectra, the current state of research looks less encouraging. The only large scale attempt to classify molecular substructures from mass spectral data was made in 1990 by Curry and Rumelhart.[14] Their discussion of several mass spectral training problems along with solution attempts provides valuable insights, but their results cannot be evaluated easily. The differences between training and test set or present and absent substructures were not used consistently throughout the paper, and the boundaries between these terms blurred frequently. However, Curry and Rumelhart found that their MSnet produced better classification results than STIRS[15] for 12 out of 16 substructures. On a less encompassing scale, Varmuza *et al.*[16] compared the ability of four chemometric methods including neural networks to classify mass spectra into 54 different structural properties. They found that neural networks performed somewhat better than *k*-nearest neighbor, linear discriminant analysis, or SIMCA methods, but attributed the generally discouraging results to the difficulty of selecting appropriate structural entities for discrimination with mass spectral data. Other approaches to neural network assisted interpretation of mass spectra tackled very specific problems, including the successful identification of 22 partially methylated alditol acetates.[17] Lohninger *et al.* employed 10 different spectral features extracted from mass spectra to classify steroids, barbituric acid derivatives, and polycyclic aromatic hydrocarbons with a >90% overall success rate.[18,19] In a very limited study, Harrington employed temperature-constrained back-propagation neural networks to successfully obtain classifiers for primary, secondary, tertiary alcohols, and ethers from mass spectra, which had been collected from compounds with the molecular formula $C_8H_{18}O$.[20] Only in the evaluation of complex pyrolysis mass spectra did neural networks seem to be more successful than other techniques.[21−23]

Reports of joint utilization of both infrared and mass spectral data have been sparse, although they complement each other well and offer distinct advantages in conjunction with separation techniques.[24] Curry[25] developed a program which pools the responses from both an IR expert system and STIRS[15] for structure elucidation, and Cooper and Wilkins showed that spectral search in a combined IR/MS library yields more reliable results than using separate libraries.[26] A combined IR/MS library was compressed with PCA by 75%, but the results were not compared with either uncompressed or separate library compression.[27] However, merging infrared and mass spectra for neural network assisted structure elucidation has never been attempted before. A publication by Gasteiger *et al.* might suggest otherwise (judging by its title "Neural Nets for Mass and Vibrational

[‡] Current address: Applied Automation, P.O. Box 9999, Bartlesville, OK 74005-9999.

[⊗] Abstract published in *Advance ACS Abstracts,* March 1, 1996.

0095-2338/96/1636-0249$12.00/0

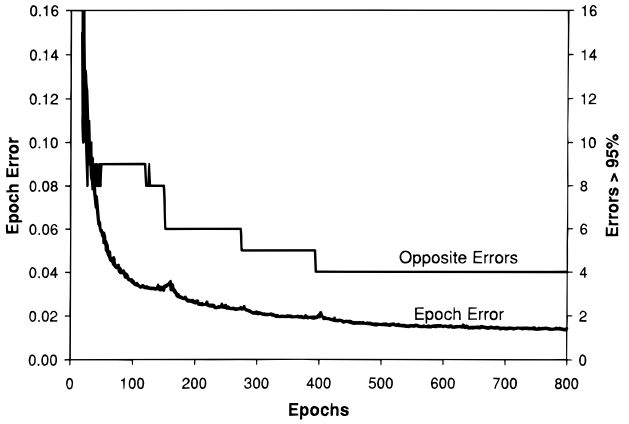

Figure 1. Error development while training 196 matrix isolation FT-IR spectra with a standard back-propagation algorithm to recognize 8 different bond types (for training details see ref 30).

Spectra"), but instead it reviewed developments for IR and MS interpretation separately.[28] Even the combination of any other spectroscopic data for neural network interpretation has not been reported yet, except for very recent work[29] which utilizes both IR and ^{13}C-NMR data. The present paper will offer some new insights into functional group interpretation with more than one spectroscopic technique, and how much infrared and mass spectroscopy may contribute to the identification of individual functional groups.

2. NEURAL NETWORK TRAINING

For this study, a heavily modified back-propagation algorithm provided the training procedure for the feed-forward neural network with a single hidden layer. The basic details of the back-propagation algorithm have been described previously,[10,30] along with the development and improvement of our modification, the "flashcard algorithm", whose concepts and algorithmic steps are summarized below.

Training multioutput neural networks to recognize molecular substructures from infrared spectra can be a time-consuming task, especially when the training set is comprised of many spectra. In addition, the natural distribution of functional groups among these spectral examples makes it impossible to compile balanced training sets, leaving some structural fragments grossly underrepresented. Because the neural network strives to find the smallest prediction error for a training set, network training is always dominated by abundant substructures such as aromatic bonds and methyl groups, making it difficult for the network to incorporate rare groups like nitriles or tertiary alcohols. Local training minima are inevitable in such a setting, and with tools provided by the back-propagation algorithm, it is often impossible to escape from these minima to a more desirable network state which also includes satisfactory prediction of less often encountered groups. The local minima manifest themselves in the appearance of "opposite errors" at the output nodes, *i.e.*, the network predicts an absent functional group with high confidence, when it is actually present, and vice versa. During training, some of these opposite errors spontaneously convert to their desired output, but they never disappear altogether. Figure 1 illustrates a typical training session, where errors >95% initially disappear as fast as the epoch error drops, but, after a while, only a few more conversions occur (indicated by the steps in the "opposite error" curve). The blips in the epoch error curve underneath the steps indicate that the conversions are indeed caused by local minimum escapes as the network moves out from one into another minimum.

Table 1. Error Types Used in the Flashcard Algorithm

error type	range	remarks
extreme	>99.5%	Prominent with too few hidden nodes or too large a training rate. Must be kept in check at all times, because they tend to gravitate toward 100.0%, after which conversion is almost impossible.
opposite	>95%	Including extreme errors, these represent the local minima, which can solidify into extreme errors.
large	L%−95%	Not a problem, because they are much closer to the "indecisive" state of the network. Their abundance may be used to gauge the training progress instead of monitoring the epoch error. L% should be <50%, usually 20−30%.
small	<L%	Desirable end state when all network outputs produce errors < L %.

In order to alleviate this problem, the flashcard algorithm was developed. It is named by analogy to learning vocabulary words using a stack of index cards: Easy words are quickly memorized and put aside, leaving a residue of difficult words which must be viewed much more often in order to be retained. Although this sounds simple enough, several checks and balances have to be implemented to keep the flashcard algorithm efficient. This is especially important when adverse training conditions such as too many or too few hidden nodes or too great a learning rate or momentum threaten the training success by causing many extreme or opposite errors (see Table 1 for a definition of the types of errors for the flashcard algorithm). The training set imbalance is taken into account by diminishing the impact of overrepresented absent functional groups on the formation of local minima. In these cases, the output node error signal must be divided by c_i (Eq 1), which describes how skewed the distribution for the functional group i is from the perfect 50/50 balance (N = number of spectra, m_i = number of occurrences of the substructure i; see ref 10 for a more detailed description):

$$f = \frac{N - m_i}{m_i} \tag{1a}$$

$$c_i = \begin{cases} f \times 2 & (f < 1) \\ f \div 2 & (f \geq 1) \end{cases} \tag{1b}$$

Multiplying the output error with c_i for present functional groups had only detrimental effects on the training progress. After these considerations, the complete flashcard algorithm proceeds as follows.

1. Calculate c_i for all i output nodes.

2. During a regular training epoch, keep track of which spectral examples cause large, opposite, and extreme errors (see Table 1).

3. After each epoch, present all spectra in the list of extreme errors to the network. As soon as a spectrum exhibits no more extreme errors, remove it from the list. Keep presenting the spectra in the list, until the list is empty. (Alternately, abort the training or resume with step 4, if the list does not shrink after $k \times N$ spectral presentations ($k \sim$

Table 2. Training Times with Different Computer Systems for a 256-25-26 Neural Network Architecture, Regular Back-Propagation Algorithm

computer/processor	operating system	time per spectrum (ms)	spectra per second
Starlab MC 68020−50[a]	Unix	6946	0.144
IBM AT, Intel 80286−8	DOS	2791	0.358
Mac II, MC 68020−16	Mac OS	817	1.22
Mac IIsi, MC 68030−33	Mac OS	451	2.22
MicroVAX II[a]	VMS	445	2.25
VAX 6310[a,b]	VMS	92.8	10.8
Intel 486DX-33	DOS	67.7	14.8
VAX 8820, 1 CPU[a,b]	VMS	67.0	14.9
Mac Centris 650, MC 68040−25	Mac OS	66.6	15.0
Coherent, Intel 486DX-33	Unix	53.4	18.7
Mac Quadra 800, MC 68040−33	Mac OS	51.6	19.4
NeXT, MC 68040−25	Unix	49.8	20.1
VAX 8820, 2 CPUs[a,b]	VMS	39.4	25.4
Intel 486DX2−66	DOS	36.2	27.6
Linux, Intel 486DX-33	Unix	34.9	28.6
NeXT, Intel 486DX2−66	Unix	21.5	46.5
Sun Sparc IPX, SunOS	Unix	16.6	60.3
Silicon Graphics Personal Iris/35[b]	Unix	13.2	75.6
Linux, Pentium-100	Unix	7.35	136
Cray YMP-EL94, 1 CPU	Unix	3.24	309
AlphaStation 400−233	VMS	2.69	372
AlphaServer 2100−275[b]	VMS	2.26	443
Cray-2, 1 CPU[b]	Unix	1.15	872
Cray-2, 4 CPUs[b]	Unix	0.609	1642
Cray C90, 1 CPU[b]	Unix	0.405	2470

[a] No floating point processor. [b] Multiuser system, run at off-peak hours.

20), because a stubborn extreme error can be caused by a faulty spectrum and/or structure assignment).

4. When the number of large errors drops below N, set c_i to 1.0. Now, present only spectra in the list of opposite errors to the network. As soon as a spectrum exhibits no more opposite errors, remove it from the list. Keep presenting the spectra in the list, until one of the following conditions is true: (a) list is empty; (b) more than $l \times N$ spectral presentations since entering step 4 ($l \sim 2$); (c) p fewer opposite errors since entering step 4 ($p \sim 5$).

5. If there are no opposite errors at the beginning of step 4, present only spectra in the list of large errors to the network. As soon as a spectrum exhibits no more large (or opposite) errors, remove it from the list. Keep presenting the spectra in the list, until one of the following conditions is true: (a) list is empty; (b) more than $l \times N$ spectral presentations since entering step 5 ($l \sim 2$); (c) q fewer opposite errors since entering step 5 ($q \sim 5$).

6. If there are any large or opposite errors left at this point, go back to step 2. Otherwise, run the entire training set one more time to check whether the network state has converged to the global minimum. Go back to step 2, if any large or opposite errors show up again.

The entire neural network software has been written in standard C and runs without changes on DOS, Mac, Unix, and VMS platforms (see Table 2 and Figure 2 for performance figures). The neural network operations described in this paper were largely carried out using a Digital AlphaStation 400 4/233 running OpenVMS 6.1-1H2.

3. DATA SELECTION

Infrared Spectra. In our previous publications,[10,30] we used the Mattson Cryolect collection of 5000 matrix isolation FT-IR spectra (ATI/Mattson, Madison, WI), but the poor quality of this database prompted us to look for alternatives. The NIST/EPA gas-phase infrared database proved to be of much higher quality and was obtained from NIST in JCAMP-DX format (Database 35, National Institute of Standards and Technology, Gaithersburg, MD). This database comes in two spectral sets, both with one data point every 4 cm^{-1}: (a) NIST file, 2120 spectra, 825 data points between 550 and 3846 cm^{-1} and (b) EPA file, 3108 spectra, 880 data points between 450 and 3966 cm^{-1}. No duplicates exist in this collection of 5228 spectra, but only 5191 spectra have structures assigned to them. All structures were checked for

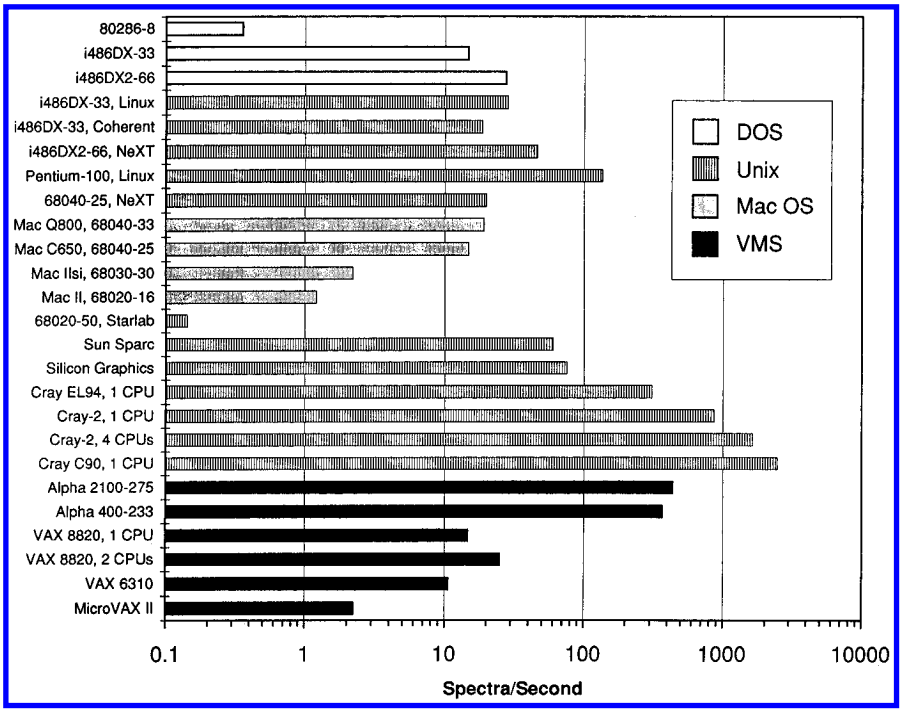

Figure 2. Training speed comparison for a 256-25-26 regular back-propagation neural network on different computers and operating systems.

Table 3. Data Reduction from 825 to 512 Data Points for the NIST/EPA Gas-Phase Infrared Database

825 data points			512 data points		
points	cm^{-1} range	Δ cm^{-1}	points	cm^{-1} range	Δ cm^{-1}
1−356	550−1970	4	1−356	550−1970	4
357−824	1974−3842	4	357−512	1982−3842	12
825	3846	4	omitted		

consistency with the associated molecular formula, and a program was written in Borland C++ 4.0 (Borland, Scotts Valley, CA) to scrutinize each structure for suspicious atom valencies, illegal bonds, and incorrect double bond equivalents. A large number of molecular formulas was found to be incorrect in the structure file, but most of these errors had been rectified in the JCAMP files. Besides these errors, a total of 24 other structures/molecular formulas were determined to be faulty and corrected subsequently. Excluding seven structures with bond uncertainties and three with more than one molecule, the available pool of spectra shrank to 5181 compounds. Although this set of molecules contained 19 different types of atoms and 66 different bond types, no further reduction was deemed to be necessary, so that the results of this study could stay as close to reality as possible.

It was necessary to generate a uniform data set with a reasonably low number of data points from the spectral files. First, all EPA spectra were truncated to the NIST range of 825 data points, followed by a reduction of the entire set of spectra to 512 points each to facilitate data compression methods such as Hadamard and Fourier transform. Table 3 lists the data points and resolution before and after the reduction procedure. This set was further reduced to 256 data points with a Hadamard transformation,[31] followed by scaling the absorbance values of every spectrum between 0.0 and 1.0. Even though 128 data points have been shown to produce satisfactory results,[32] this could have created an information disadvantage for IR spectra in comparison with the more than 400 data points for mass spectra (see below).

Mass Spectra. The NIST/EPA/NIH mass spectral database (CD-ROM, ASCII version) served as the source for MS data (National Institute of Standards and Technology, Gaithersburg, MD), and fortunately, all 5181 IR counterparts were found in the MS data base, so that no further structure manipulation was necessary. Adequate representation of mass spectral data turned out to be a much more difficult problem, because neural networks require a fixed-length data input vector, but mass spectra are "open-ended" by nature. This obstacle is compounded by the fact that, in many cases, structural information in mass spectra cannot simply be deduced by the presence or absence of particular peaks in the spectrum. Peak differences, fragmentation patterns, and present or absent molecular ion peaks play an equally important role in structure elucidation. A good starting point in this matter was the paper by Curry and Rumelhart,[14] who used not only mass-to-charge data points but also neutral loss spectra and a number of correlation methods to augment the primary mass spectrum. The mass range below 40 was excluded because it frequently contains highly abundant peaks of limited diagnostic value, and library spectra often do not include this range at all. The need for accurate molecular weight information may constitute a problem when using neutral loss spectra to identify unknown spectra, but this restriction is less serious because a lot of work has gone into the estimation of this molecular property from mass spectra.[33−38]

Because several modifications to the equations described by Curry and Rumelhart were deemed necessary, they will be reproduced here. Prior to any data treatment, the entire spectrum was scaled between 0.0 and the base peak abundance $B = 1000.0$. Table 4 summarizes the features and formulas used in this paper and includes the definitions for the feature symbols in this paragraph. Logarithmic scaling of S_i between 0.0 and 1.0 causes minor peaks to gain in importance relative to the base peak, and neutral loss peaks N_i make fragmentation patterns accessible which do not

Table 4. Mass Spectral Features and Data Processing[a]

#	data points	feature	processing
A	180	mass spectrum, $i = 40-219$	$S_i = \begin{cases} 0.0 & I_i \le 1 \\ \frac{1}{3}\log_{10}(I_i) & I_i > 1 \end{cases}$
B	180	neutral loss spectrum, $i = 0-179$	$N_i = S_{m-i}$
C	50	autocorrelation sums, $k = 1-50$	$A_k = \sum_{i=1}^{m} (I_i I_{(i+k-1) \bmod k+1})/A_0$ $A_0 = \sum_{i=1}^{m} (I_i)^2$
D	50	constants sums, $k = 0-49$	$C_k = \sum_{i=1}^{m} (I_k I_{m+k-i})/10^6$
E	14	*mod*-14 series sums, $k = 0-13$	$x = \begin{cases} 11 & (k+m) \bmod 14 \ge 11 \\ (k+m) \bmod 14 & (k+m) \bmod 14 < 11 \end{cases}$ $K_k = \frac{1}{x}\sum_{i=1}^{x} S_{k+14i}$
F	14	*mod*-14 neutral loss series sums	calculated similarly to K_k, except from neutral loss spectrum instead of mass spectrum

[a] I_i = intensity of m/z i after scaling between 0.0 and 1.0; m = mass of molecular weight ion.

Table 5. Classifier Abundance in the Database: Test and Training Sets

functional group	database	test set	training set 1	training set 2	training set 3
O−H	1264	214	221	630	644
N−H	599	111	101	297	302
C−N	1089	200	201	533	545
X≡Y, X═Y═Z	326	44	44	163	163
C═O	1664	203	201	599	832
C−O	2403	276	299	874	1202
C═C	952	67	73	202	483
C:::C	2352	259	264	783	1189
halogen	1193	84	77	275	597
N−O, N═O	260	44	44	130	130
CH_3, CH_2	4392	410	420	1282	2196
1° alcohol	297	44	44	149	149
2° alcohol	305	44	44	153	153
3° alcohol	143	44	44	72	72
Ar−OH	302	44	44	151	151
COOH	228	44	44	114	114
1° amine	286	44	45	143	149
2° amine	187	44	44	94	94
3° amine	157	44	44	79	79
(CO)NH	249	44	44	126	125
(CO)OR	671	44	49	171	338
(CO)H	125	44	44	63	63
C(CO)C	460	44	44	171	233
arom C_6-ring	2231	230	238	725	1132
pyridine ring	184	44	44	92	92
C−Cl	680	50	61	171	343
total spectra	5181	500	511	1525	2588

produce ions in a mass spectrometer. Autocorrelation sums reveal repeating distances between peaks, *e.g.*, aliphatic hydrocarbons display a large value at A_{14}. However, the algorithm in ref 14 resulted in some abnormally high outliers, possibly because the mass spectrum was not viewed as a "circular" set of data. The equation in Table 4 takes this into account.[31] Constant sums reveal symmetrical fragmentation patterns in the spectrum. For example, ketones may fragment on either side of the carbonyl group to give a characteristic constant sum for C_{28}. Unfortunately, Curry and Rumelhart's equation did not produce correctly scaled values, because normalization was carried out with the base peak instead of the square of the base peak ($= 10^6$). This has been corrected as well. On a few occasions, peak-rich spectra cause certain C_k to become >1. In these cases, all C_k for this spectrum were scaled back to values between 0.0 and 1.0. *Mod*-14 series have been shown to be a good measure for spectral similarity,[39] especially for aliphatic hydrocarbons, making them and their neutral loss complement a good addition to the feature list in Table 4. Only the "odd series" sums were not included from the feature list in ref 14. As we will see in section 4, not all of the features necessarily contribute to better predictions of functional groups from mass spectra.

Training and Test Sets. Realization of a fair distribution of present and absent functional groups within the training and test sets turns out to be a difficult task, because their natural distribution varies strongly. This task becomes harder with an increasing number of spectra in a training or test set. In this work, a set of 26 functional groups or classifiers (Table 5) had to be shaped into training and test sets. These functional groups were selected because of their distinct infrared absorbance patterns and their abundance in the NIST spectral databases. For training sets with fewer spectra, an algorithm was devised which allows a fairly even distribution among low abundance classifiers, using *N*% as a definition

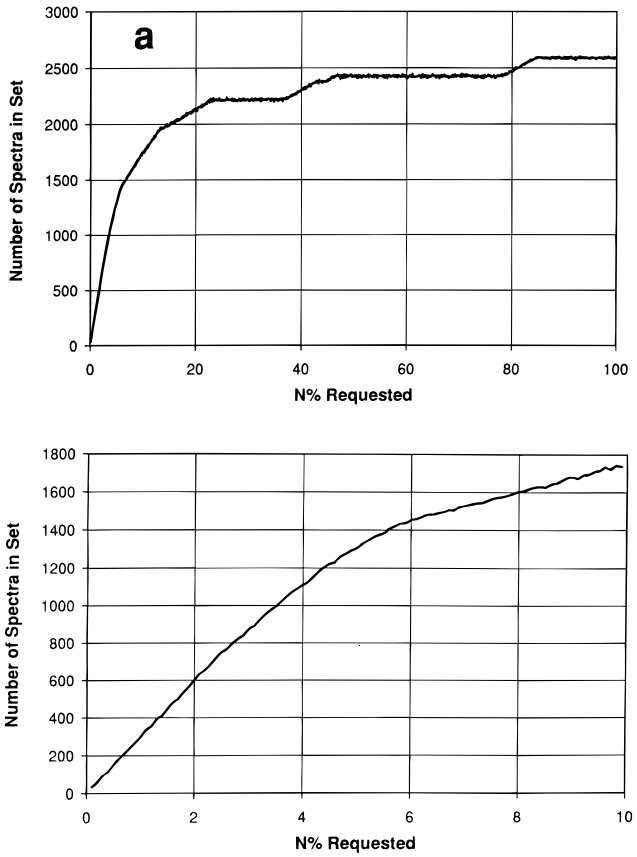

Figure 3. Correlation between the *N*% definition of "low abundance" classifiers and the resulting number of spectra in a training set: (a) full curve and (b) enlarged section for small *N*%.

of "low abundance". Varying *N*% between 0 and 100 in the algorithm described below, the resulting number of spectra in a training set grew as illustrated in Figure 3. Once this graph was set up, it was easy to predict the necessary *N*% for a desired number of spectra, and then run the program again to compile a classifier set. For example, a training set with approximately 500 spectra requires $N\% \approx 1.7$, and for ~1520 spectra, $N\% \approx 6.9$.

1. Determine the minimum abundance F_i of a classifier in the training set (T_i is the total classifier count in the database, D the number of spectra in the database)

$$F_i = \begin{cases} 0.5 \times T_i & 0.5 \times T_i < N\% \\ 0.5 \times N\% \times D & 0.5 \times T_i \geq N\% \end{cases}$$

2. Sort F_i for all i in ascending order. If two classifiers have the same F_i, rank the one with the smaller T_i first.

3. Randomize the order of the pattern list, which contains the presence/absence of each classifier for every spectrum.

4. Repeat step 4 for all F_i, lowest one first, starting at the top of the pattern list each time (keep in mind that each pattern is often comprised of more than one classifier): (a) Pick the next pattern containing F_i from the pattern list. (b) Increase the count C_n for *all* present classifiers in this pattern. (c) Move this pattern from the pattern list to the training set list. (d) Repeat (a)−(c) until $C_i \geq F_i$.

This algorithm was used to produce three training sets with 511, 1525, and 2588 spectra, and a test set encompassing 500 spectra, which does not overlap with any of the training sets.

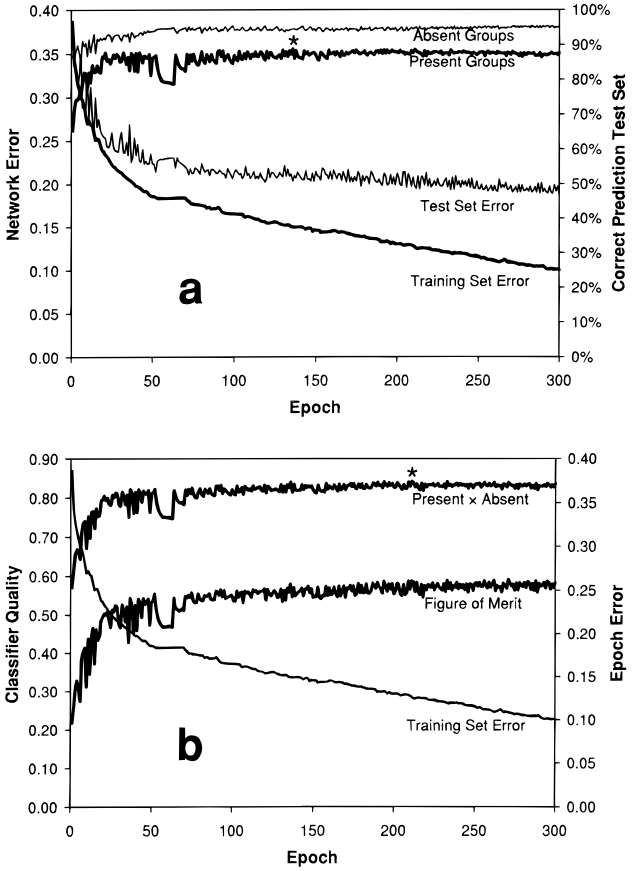

Figure 4. Training of 1525 infrared spectra to recognize 26 different functional groups, concurrent testing during training with 500 spectra. (a) Training and test set errors during training with correct predictions of present and absent functional group in the test set. The asterisk (*) marks the maximum of correctly predicted present functional groups (89.1%). (b) Progress of the network quality as a classifier. The asterisk shows the optimum prediction for combined present × absent classifiers.

4. SEPARATE TRAINING

Infrared Spectra. A training set of 1525 spectra seemed to provide a broad enough scope for the neural network to learn spectra-structure correlations in a satisfactory fashion. During the training of the neural network with 256 input nodes, 50 hidden, and 26 output nodes, all spectra of the test set were presented to the network after each epoch to gauge the progress in the generalization ability of the network (1% constant input noise during training, learning rate $\eta = 0.02$, momentum $\alpha = 0.9$, sigmoidal function steepness $\beta = 1.0$, sigmoidal x-axis shift $\theta = 0.0$). However, merely measuring the overall output error for training and test set does not paint an accurate picture of the network's ability to correctly predict present and absent functional groups in the test set. Although the training and test set error curves in Figure 4a continue to drop, the prediction quality for present functional groups reaches a maximum at 136 epochs (marked with an asterisk), before starting to decrease very gradually. On the other hand, recognition of absent groups keeps rising further, raising the question of a reliable measure of what constitutes the "best" prediction state. Wilkins and co-workers[40] proposed a figure of merit which removes the dependency on the test set composition from the prediction figures. In essence, it tells how much better a classifier can predict substructures compared with random guessing. For a more in-depth discussion of this figure of merit see ref 40.

Table 6. Suitability of Mass Spectral Features (see Table 4 for Description) for Neural Network Training and Prediction with 511 Training and 500 Test Spectra (25 Hidden Nodes Used in All Cases)

used features	present × absent[a]	epochs[b]
All	0.6146 ± 0.0022	19.2
no A	0.5949 ± 0.0028	20.6
no B	0.5844 ± 0.0029	35.6
no C	0.6146 ± 0.0007	23.2
no D	0.6200 ± 0.0018	22.6
no E	0.6174 ± 0.0019	22.8
no F	0.6160 ± 0.0006	19.0
A, B	0.6116 ± 0.0035	18.0
A, B, C	0.6155 ± 0.0010	29.8
A, B, D	0.6114 ± 0.0014	18.6
A, B, E	0.6194 ± 0.0016	18.0
A, B, F	0.6147 ± 0.0023	18.0
A, B, C, E	0.6196 ± 0.0018	25.4

[a] Average best results for fractions of correctly predicted present groups × absent groups. [b] Average training epochs needed to reach best prediction point.

However, this does not alleviate the fact that present functional groups are at a severe disadvantage in the *training* set, and the figure of merit continues to rise in parallel with correctly predicted absent groups (Figure 4b).

Therefore, a different approach, which judges how well a classifier can separate present from absent predictions, is proposed here. Multiplying the fractions of correctly found present and absent groups will only yield a high value when *both* present and absent groups are predicted well. This means that when a lot of effort is spent improving the prediction quality of present functional groups, while at the same time the classifier has difficulties with absent groups, nothing has been gained. In this case, easing the requirements for present groups gained more for absent groups to reach a combined maximum after 211 epochs (marked with an asterisk in Figure 4b). In this particular example, the advantage is rather small but becomes much more important when the network is trying to locate a less distinct training minimum. Overall, the network was able to correctly predict 88.4% of all present and 95.5% of all absent functional groups.

Mass Spectra. Which combinations of the six mass spectral features in Table 4 would yield a neural network with the best prediction power? This question had to be answered first with a series of tests using a training set of 511 spectra (same training parameters as for IR spectral training). Systematically omitting one of the features from the training set, the biggest drop in the prediction capability was encountered when the neural network was trained without the actual mass spectral peaks or the neutral loss series. Comparing the numbers in the prediction quality column "present × absent" in Table 6, it appears that the neutral loss series carries the most information, with the actual mass spectral peaks being second in information content. Some features like the constant sum (D) even appear to lead to confusion, because their omission causes the prediction rate to go *up* compared with using all features together.

From Table 6, it would follow that the combination of mass spectral peaks, neutral loss spectrum, and *mod*-14 series (A, B, and E) yields the best prediction with the fewest number of epochs. However, when the training set was expanded to 1525 spectra and the neural network to 50

Table 7. Percentage of Correctly Predicted Present Functional Groups with Separate and Combined Training of IR and Mass Spectra and Using Only IR or MS Input Data in a Combined IR/MS Network for Prediction[a]

functional group	random guessing	IR	MS	IR/MS	IR/MS IR input	IR/MS MS input
O−H	42.8%	**93.5%**	71.0%	84.6%	72.0%	52.8%
N−H	22.2%	**77.5%**	68.5%	74.8%	66.7%	27.9%
C−N	40.0%	85.0%	83.5%	**88.0%**	70.0%	75.5%
X≡Y, X=Y=Z	8.8%	72.7%	52.3%	**75.0%**	20.5%	56.8%
C=O	40.6%	**98.5%**	76.4%	96.6%	97.5%	40.4%
C−O	55.2%	**95.3%**	88.8%	92.8%	93.1%	49.6%
C=C	13.4%	**71.6%**	53.7%	64.2%	37.3%	17.9%
C:::C	51.8%	**96.9%**	91.9%	93.1%	87.3%	64.1%
halogen	16.8%	59.5%	73.8%	**79.8%**	34.5%	79.8%
N−O, N=O	8.8%	**88.6%**	65.9%	**88.6%**	84.1%	18.2%
CH_3, CH_2	82.0%	**97.6%**	95.1%	95.9%	96.3%	91.0%
1° alcohol	8.8%	**77.3%**	65.9%	70.5%	47.7%	6.8%
2° alcohol	8.8%	**77.3%**	70.5%	**77.3%**	31.8%	13.6%
3° alcohol	8.8%	**72.7%**	59.1%	68.2%	20.5%	13.6%
Ar−OH	8.8%	**95.5%**	50.0%	70.5%	52.3%	0.0%
COOH	8.8%	93.2%	81.8%	**97.7%**	75.0%	4.6%
1° amine	8.8%	79.6%	63.6%	**84.1%**	56.8%	11.4%
2° amine	8.8%	**75.0%**	50.0%	56.8%	22.7%	9.1%
3° amine	8.8%	**75.0%**	70.5%	70.5%	27.3%	6.8%
(CO)NH	8.8%	81.8%	70.5%	**86.4%**	61.4%	6.8%
(CO)OR	8.8%	**97.7%**	68.2%	90.9%	81.8%	2.3%
(CO)H	8.8%	**86.4%**	52.3%	77.3%	34.1%	18.2%
C(CO)C	8.8%	**72.7%**	54.6%	63.6%	68.2%	2.3%
arom C_6-ring	46.0%	**95.2%**	85.7%	90.9%	84.4%	58.7%
pyridine ring	8.8%	65.9%	52.3%	**68.2%**	52.3%	2.3%
C−Cl	10.0%	44.0%	68.0%	**76.0%**	10.0%	74.0%
overall	21.3%	88.4%	78.2%	86.4%	73.1%	49.9%

[a] Boldface data are the best prediction rates for a functional group.

hidden nodes, the feature combination A, B, and E gave poorer prediction quality (0.6680) than employing all six features (0.6760). This could be attributed to the need of the network to make finer distinctions between the functional groups as the training set and network size grows. Compared with the IR prediction rates, mass spectra results were worse, with an overall success rate of 78.2% correctly predicted present and 87.1% absent functional groups. Contrary to common neural network training wisdom, the prediction rates could not be improved by using more training spectra: 2588 training spectra with 70 hidden nodes produced a lower prediction quality of 0.6665, and even training with 10% random input noise instead of 1% (simulating the higher noise level of mass spectra[14]) led to a drop in the prediction quality. Evidently, the current choice of molecular substructures, which were tailored specifically to the interpretation of infrared spectra, does not agree too well with mass spectra. This notion is supported by the data in Table 7, where only the halogens and the C−Cl bond generated a better prediction rate with mass spectra than with IR spectra. Incidentally, these groups produce either fairly obscured or absent bands in IR spectroscopy but can often be clearly recognized in mass spectra.

5. COMBINING IR AND MS DATA

Joint Interpretation. Merging both spectral data sets for combined training and testing simply required the allocation of 256 input nodes to IR data and 488 input nodes to MS data, which resulted in a 744-50-26 network (same number of hidden and output nodes and training parameters as used for separate spectral training runs). The joint network trained surprisingly fast and was able to perfectly recall the entire training set of 1525 spectra within ∼75 epochs. This is even more astonishing when compared with the epoch error progression of the individual IR and MS training runs (see Figure 5). MS training alone takes ∼140 epochs to reach the point of 100% correct recall of training spectra, but with IR spectra, training is even slower and perfect recall is not reached within 500 epochs. Apparently, IR spectra provide more low intensity data which are hard to incorporate into a neural network. On the other hand, mass spectra may have more abundant ions leading to larger peaks to support the presence of functional groups (especially with logarithmic scaling) and to facilitate faster learning, but the same peaks may be more ambiguous when it comes to functional group prediction. The even faster combined learning can be attributed to the higher network capacity (IR/MS: 38 500 weights; IR: 14 100 weights; MS: 25 700 weights) and to more data supporting the same information, making the

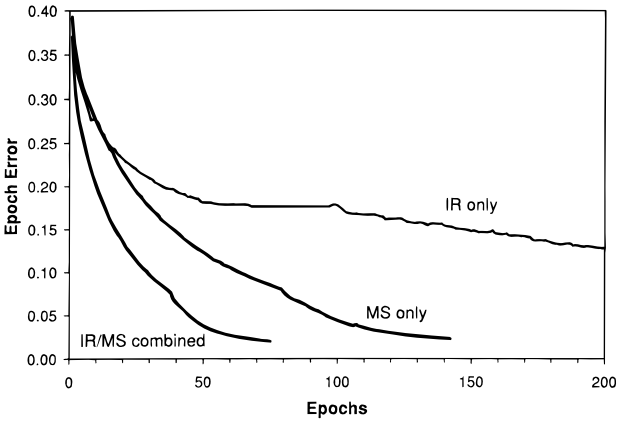

Figure 5. Epoch error development for separate and combined IR and MS training. The endpoint of 100% correct recall has been reached only for MS and IR/MS training.

network pick the best inputs from both spectroscopic methods to arrive at the same "conclusions".

Overall, the combined use of IR and MS data did not generate better prediction rates. Comparing the IR/MS column in Table 7 with the IR and MS columns, the IR performance is consistently better in most cases. However, in 8 out of 26 cases, the additional data resulted in improved prediction, again most notably for halogens and C−Cl bonds. Several nitrogen-containing groups (C−N, C≡N, primary amine, amide, and pyridine ring) also show some improvement, where the MS data might be capable of removing some ambiguity left by the infrared spectrum. But especially groups containing the IR-prominent C═O, C−O, and O−H bonds exhibited much worse results in the combined IR/MS network than using infrared spectra alone.

Partial Data Input. Two interesting questions arise at this point: How much information does the neural network draw from each of the individual spectroscopic techniques, and what happens when only IR or MS data are available for predictive purposes, but not both? This case occurs not too infrequently in hyphenated techniques, *e.g.*, gas chromatographic separation followed by IR and MS detection, where low analyte concentrations sometimes allow only the collection of mass spectra. Both questions can be answered by presenting only IR or only MS data to a network trained with IR and MS data, and setting the inputs of the other data type to zero. As it turns out, the neural network cannot operate well in either case. Using just IR inputs, the rate of correctly identified present functional groups in the test set dropped from 86.4% to 73.1%, and with MS inputs, the rate fell even further to 49.9% (approximately chance prediction), supporting the assumption that although there are fewer IR than MS data points, infrared spectra are much better suited to predict the current set of functional groups.

Several observations can be made here: Not surprisingly, halogens and the C−Cl bond draw most of their information from mass spectra, judging from the almost unchanged prediction rates compared with IR/MS data, while predictions from infrared data are completely unreliable. On the other hand, the very characteristic C═O and C−O bands in IR spectra carry so much weight for the neural network, that using only IR data actually results in a slight prediction *improvement*, but MS data do not carry any more significance than mere random guessing for these cases. The aromatic bond and C_6-ring results suggest that the current representation of MS data may not be ideal. Despite their distinct patterns in mass spectra, the neural network assigns much more significance to the IR data, although aromatic bonds often give rise to only weak bands. In general, more sophisticated substructure definitions have a better chance of being recognized with IR than MS data.

Including the correct prediction of absent functional groups into the analysis, the results look somewhat brighter. Combined prediction of IR and MS data yields approximately the same rate as IR training and prediction alone (Table 8, Figure 6), and even presenting only IR or only MS data to the combined network does not result in much lower prediction rates for absent functional groups.

6. CONCLUSIONS

The problem of balanced training and test sets for multidimensional classifiers has been addressed successfully in this paper with the development of an algorithm which allows a compromise between data set size and even distribution of individual classifiers. With this algorithm and a unified presentation of the previously developed flashcard algorithm,[10,30] infrared and mass were trained in separate and joint neural networks. The best overall prediction rates for present functional groups could be achieved using only IR data for training and testing, but for 8 of the 26 molecular substructures were predicted better using a combination of infrared and mass spectra. Thus, individual networks for selected functional groups may be constructed to take full advantage of either data constellation, using IR and MS data for these eight groups, but only IR data for the remaining 18 functionalities. The disappointing prediction results involving mass spectral data may be attributed not only to the infrared-biased selection of substructures, but also to the general difficulty of associating the more distributed mass spectral features with particular functional groups.[16]

The exceptionally rapid neural network training with MS and IR/MS data to the point of 100% correct recall of the training spectra adds to the attractiveness of a previous proposal that neural networks be employed for rapid spectral database searches.[10] Using a trained IR/MS network, the input of only IR or MS data provided some insight into the suitability of either spectroscopic method for neural network applications. In most cases, the neural network correlates the infrared data much more strongly with the functional groups, even when both IR and MS should provide evenly strong evidence for the presence of a particular functionality. Only for halogens, the addition of mass spectral data enabled the neural network to produce better classifiers, relying almost exclusively on MS data. For six other functionalities, some improvement could be observed with the joint IR/MS network. Thus, the development of molecular substructures which are much more amenable to neural network training with mass spectra should yield better predictions not only

Table 8. Overall Rates for Correctly Predicted Present and Absent Functional Groups

training data	test data	present (%)	absent (%)
IR	IR	88.4	95.5
MS	MS	78.2	87.1
IR/MS	IR/MS	86.4	95.6
IR/MS	IR only	73.1	96.0
IR/MS	MS only	49.9	94.8
random guessing		21.3	78.7

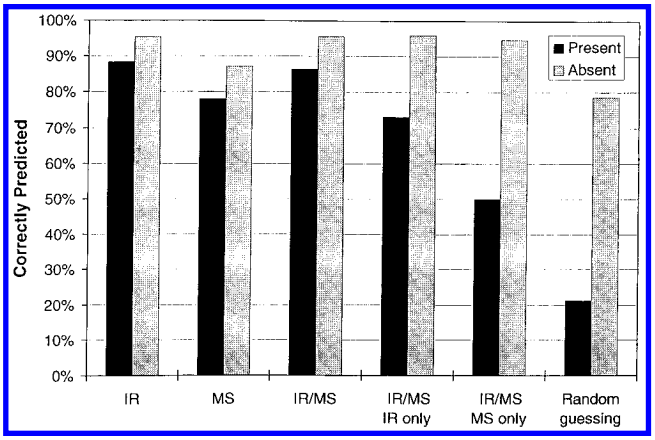

Figure 6. Overall prediction rates for correctly predicted present (black) and absent (gray) functional groups.

for mass spectra alone but also possibly for the combined interpretation of both infrared and mass spectra.

ACKNOWLEDGMENT

This work has been carried out with support from the National Science Foundation, Grant CHE-92-01277. Preliminary findings were presented at the PittCon'95 Conference in New Orleans, LA, March 1995, paper 713. The authors are indebted to Stephen Stein at the National Institute of Standards and Technology for providing us with the IR gas phase and MS spectral libraries.

REFERENCES AND NOTES

(1) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing−Explorations in the Microstructure of Cognition.* The MIT Press: Cambridge, MA, 1986; Vol. 1, pp 318−362.
(2) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533−536.
(3) Robb, E. W.; Munk, M. E. A neural network approach to infrared spectrum interpretation. *Mikrochim. Acta [Wien]* **1990**, *I*, 131−155.
(4) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural network models for infrared spectrum interpretation. *Mikrochim. Acta [Wien]* **1991**, *II*, 505−514.
(5) Weigel, U.-M.; Herges, R. Automatic interpretation of infrared spectra: Recognition of aromatic substitution patterns using neural networks. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 723−731.
(6) Meyer, M.; Weigelt, T. Interpretation of infrared spectra by artificial neural networks. *Anal. Chim. Acta* **1992**, *265*, 183−190.
(7) Ricard, D.; Cachet, C.; Cabrol-Bass, D. Neural network approach to structure feature recognition from infrared spectra. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202−210.
(8) Tanabe, K.; Tamura, T.; Uesaka, H. Neural network system for the identification of infrared spectra. *Anal. Sci.* **1991**, *7*, 717−718.
(9) Tanabe, K.; Tamura, T.; Uesaka, H. Neural network system for the identification of infrared spectra. *Appl. Spectrosc.* **1992**, *46*, 807−810.
(10) Klawun, C.; Wilkins, C. L. Neural network assisted rapid screening of large infrared spectral databases. *Anal. Chem.* **1995**, *67*, 374−378.
(11) Fessenden, R. J.; Györgyi, L. Identifying functional groups in IR spectra using an artificial neural network. *J. Chem. Soc., Perkin Trans. 2* **1991**, 1755−1762.
(12) Van Est, Q. C.; Schoenmakers, P. J.; Smits, J. R. M.; Nijssen, W. P. M. Practical implementation of neural networks for the interpretation of infrared spectra. *Vib. Spectrosc.* **1993**, *4*, 263−272.
(13) Smits, J. R. M.; Schoenmakers, P.; Stehmann, A.; Sijstermans, F.; Kateman, G. Interpretation of infrared spectra with modular neural-network systems. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 27−39.
(14) Curry, B.; Rumelhart, D. E. MSnet: A neural network which classifies mass spectra. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213−237.
(15) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. Computer-aided interpretation of mass spectra. III. A self-training interpretive and retrieval system. *J. Am. Chem. Soc.* **1973**, *95*, 4185−7194.
(16) Werther, W.; Lohninger, H.; Stancl, F.; Varmuza, K. Classification of mass spectra. A comparison of yes/no classification methods for the recognition of simple structural properties. *Chemom. Intell. Lab. Syst.* **1994**, *22*, 63−76.
(17) Sellers, J.; York, W.; Albersheim, P.; Darvill, A.; Meyer, B. Identification of the mass spectra of partially methylated alditol acetates by artificial neural networks. *Carbohydrate Res.* **1990**, *207*, c1−c5.
(18) Lohninger, H. Classification of mass spectral data using neural networks. *Software Development in Chemistry 5*; Gmehling, J., Ed.; Springer-Verlag: Berlin, 1991; pp 159−168.
(19) Lohninger, H.; Stancl, F. Comparing the performance of neural networks to well-established methods of multivariate data analysis: The classification of mass spectral data. *Fresenius J. Anal. Chem.* **1992**, *344*, 186−189.
(20) Harrington, P. B. Temperature-constrained backpropagation neural networks. *Anal. Chem.* **1994**, *66*, 802−807.
(21) Goodacre, R.; Kell, D. B.; Bianchi, G. Rapid assessment of the adulteration of virgin olive oils by other seed oils using pyrolysis mass spectrometry and artificial neural networks. *J. Sci. Food Agric.* **1993**, *63*, 297−307.
(22) Chun, J.; Atalan, E.; Kim, S.-B.; Kim, H.-J.; Hamid, M. E.; Trujillo, M. E.; Magee, J. G.; Manfio, G. P.; Ward, A. C.; Goodfellow, M. Rapid identification of streptomycetes by artificial neural network analysis of pyrolysis mass spectra. *FEMS Microbiol. Lett.* **1993**, *114*, 115−120.
(23) Goodacre, R.; Neal, M. J.; Kell, D. B. Rapid and quantitative analysis of the pyrolysis mass spectra of complex binary and tertiary mixtures using multivariate calibration and artificial neural networks. *Anal. Chem.* **1994**, *66*, 1070−1085.
(24) Wilkins, C. L. Linked GC/IR/MS *Anal. Chem.* **1987**, *59*, 571A−581A.
(25) Curry, B. An expert system for organic structure determination. *Artificial Intelligence Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 350−364.
(26) Cooper, J. R.; Wilkins, C. L. Utilization of spectrometric information in linked gas chromatograph−Fourier transform infrared spectroscopy−mass spectrometry. *Anal. Chem.* **1989**, *61*, 1571−1577.
(27) Williams, S. S.; Lam, R. B.; Isenhour, T. L. Search system for infrared and mass spectra by factor analysis and eigenvector projection. *Anal. Chem.* **1983**, *55*, 1117−1121.
(28) Gasteiger, J.; Li, X.; Simon, V.; Novič, M.; Zupan, J. Neural nets for mass and vibrational spectra. *J. Mol. Struct.* **1993**, *292*, 141−160.
(29) Munk, M. E.; Madison, M. S.; Robb, E. W. The neural network as a tool for multispectral interpretation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 231−238.
(30) Klawun, C.; Wilkins, C. L. A novel algorithm for local minimum escape in back-propagation neural networks: Application to the interpretation of matrix isolation infrared spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 984−993.
(31) Zupan, J. *Algorithms for Chemists*; John Wiley & Sons: Chichester, 1989.
(32) Novič, M.; Zupan, J. Investigation of infrared spectra−structure correlation using Kohonen and counterpropagation neural networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.
(33) Scott, D. R. Expert system for estimates of molecular weights of volatile organic compounds from low-resolution mass spectra. *Anal. Chim. Acta* **1991**, *246*, 391−403.
(34) Scott, D. R. Improved method for estimating molecular weights of volatile organic compounds from low resolution mass spectra. *Chemom. Intell. Lab. Syst.* **1991**, *12*, 189−200.
(35) Scott, D. R. Rapid and accurate method for estimating molecular weights of organic compounds from low resolution mass spectra. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 193−202.
(36) Scott, D. R.; Levitsky, A.; Stein, S. E. Large scale evaluation of a pattern recognition/ expert system for mass spectral molecular weight estimation. *Anal. Chim. Acta* **1993**, *278*, 137−147.
(37) Scott, D. R. Empirical pattern recognition/expert system for molecular weight estimation of low resolution mass spectra. *Anal. Chim. Acta* **1994**, *285*, 209−222.
(38) Scott, D. R. Pattern recognition/expert system for identification of toxic compounds from low resolution mass spectra. *Chemom. Intell. Lab. Syst.* **1994**, *23*, 351−364.
(39) Hertz, H. S.; Hites, R. A.; Biemann, K. Identification of mass spectra by computer searching a file of known spectra. *Anal. Chem.* **1971**, *43*, 681−691.
(40) Soltzberg, L. J.; Wilkins, C. L.; Kaberline, S. L.; Lam, T. F.; Brunner, T. R. Evaluation and comparison of pattern classifiers for chemical applications. *J. Am. Chem. Soc.* **1976**, *98*, 7139−7144.

CI9501002