18 Infrared Spectrum-Structure Correlation

learning objectives:

- the classification of objects simultaneously into several classes, or hierarchy of classes
- different means of spectra representation by reduced sets of intensities, or by reduced sets of Fourier and Hadamard coefficients,
- the possibility of using different spectrum representations for different spectral regions (i.e., different functional groups)
- use of statistical methods to assist in the interpretation of Kohonen maps
- expanding a study from a Kohonen network to a counter-propagation network
- a mathematical transformation of the 3D structure of a molecule into a fixed length representation
- different ways of selecting a training set
- obtaining a 3D structure from an infrared spectrum

18.1 The Problem

Previously, we have seen classification problems where an object has to be assigned to **one** of several categories. Now, we will look at an example where the object has to be assigned **simultaneously to several** classes out of many possible ones.

After assigning compounds to various structure classes on the basis of their infrared spectra, a modeling of the relationships between structure and infrared spectra is presented that leads to the simulation of high quality infrared spectra. By carrying this work further it will be demonstrated how the 3D structure of a molecule can be derived from its infrared spectrum.

The objects of our present example are the infrared spectra of various compounds. The output of the neural network should be a series of substructures that are contained in the compound whose infrared spectrum is being investigated. Except for the immediately preceding example (Chapter 17 – the secondary structure of proteins), our previous examples have had rather simple neural networks with small numbers of weights; in this application, we will meet much larger multilayer neural networks containing 10,000 or more weights.

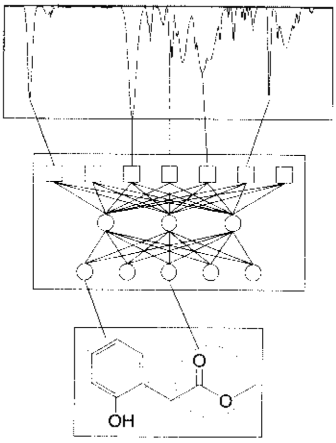


Figure 18-1: The problem: derivation of substructural features from the infrared spectrum.

The elucidation of the structure of organic compounds relies heavily on spectroscopic methods. However, the relationships between structure and spectral data are usually too complex to be expressed as explicit equations. As in many complex associational problems (medical diagnosis, for example), a series of empirical rules has been developed. The search for structure/spectra methods can, fortunately, build on a host of experimental data, much of which is now available in computerized databases.

Today, along with high-resolution full-curve spectrum, the databases also contain chemical structure coded as a connection table. The clear (though complicated!) relationships between structure and spectra, and the availability of large computerized datasets (50,000 spectra now, and more every year) make this field an ideal – and important – area of application for neural networks.

The work presented in this chapter stresses the importance of structure representation. Both, the investigations reported in Sections 18.2 - 18.3 and those in Sections 18.4 - 18.6 represent structures by a set of functional groups. In these studies, the objective is to predict the presence or absence of functional groups from the information contained in an infrared spectrum. The two different studies allow a comparison of what can be achieved with a back-propagation network against results from a Kohonen network. On the other hand, a representation of structures by functional groups is hopelessly inadequate for the reverse problem, for the simulation of an infrared spectrum over the entire frequency domain. A break-through in this area could only be achieved by the use of a novel molecular transform of the 3D structure. This structure representation then allowed even the derivation of the 3D structure from the information present in an infrared spectrum (Figure 18-1).

This chapter only deals with the application of neural networks to structure-infrared spectrum correlations. However, work on finding correlations between structure and data from other spectroscopic methods like **mass spectra** or **^{13}C NMR spectra** with neural networks has already been done. Some of these investigations are mentioned in Section 18.7 (References).

18.2 The Representation of Infrared Spectra as Intensities

Munk, Madison, and Robb, (Reference 18-1) represented infrared spectra in the following way. As a first approach, the range of a spectrum from 4000 – 400 cm^{-1} was divided into 640 intervals of width 5.6 cm^{-1}. The transmission intensity value of one interval was then scaled according to the equation:

$$x_i = 1.00 - (\%t) / 100.0 \qquad (18.1)$$

where $\%t$ = % transmission

Thus the neural network would need 640 input units; but this large number of input units caused some spurious results, and it was reduced to 256. At the same time, they adjusted the widths of the intervals to be narrowest at low frequencies and broadest at the high frequency end of the spectrum (to take into account the varying discrimination from one end of the spectrum to the other). The formula that makes the **length** of the interval i dependent on the frequency is:

$$i = 6.0\,(\mathit{frequency})^{0.5} - 120.0 \qquad (18.2)$$

rounded to the nearest integer. This equation assigns a frequency interval of 10 cm^{-1} (from 400 – 410 cm^{-1}) to input unit 1; on the other end of the spectrum, it assigns a frequency interval of 20 cm^{-1} (from 3928 – 3948 cm^{-1}) to input unit 256. The assigned frequency interval for each unit is then scanned for peaks; if a peak is found, its intensity (scaled to lie between 0.000 and 1.000) is the input to this unit, otherwise the input to the unit is zero.

The structure of the compound is described in terms of 36 functional groups (primary alcohol, phenol, tertiary amine, ester, etc.), each represented by one output unit. Hence, the target vector is a 36-

variate binary vector in which each 1 indicates the presence of the associated functional group, and zero indicates its absence.

In general, a structure can have several such functional groups, and thus several output units might be simultaneously active. After trying 14 different networks varying from fewer than ten to more than 60 neurons in the hidden layer, 34 were found to be appropriate. Thus, they used a neural network having just under 10,000 ((256 + 1) x 34 + (34 + 1) x 36 = 9,998) weights. Figure 18-2 shows the network used in this example.

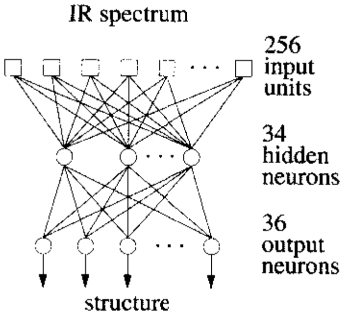


Figure 18-2: The network for the infrared spectra-structure correlation problem.

18.3 The Dataset, and Learning by Back-Propagation

If the back-propagation algorithm is used to train a large neural network such as the one described in the previous section, we must have a large training set. A good rule of thumb is that the number of data values taken for training should be equal to or greater than the number of weights to be determined in the network. Here, the data matrix contains about 640,000 values (2,499 spectra times 256 intensities), which is about 60 times larger than the number of weights. An additional 416 spectra were set aside to test the prediction ability of the trained neural network.

A relatively small learning rate η of 0.083 was used. One epoch of training using the entire dataset of 2500 spectra required about 10 min of CPU time on a VAX 3500; typically 100 epochs were needed to stabilize the network.

The actual outputs of the network are seldom exactly zero or one. Therefore, predicting the presence or absence of a functional group strongly depends on where the threshold is set. For example, in this network only 30% of the 265 primary alcohols contained in the data set produce an output equal to 1.0. However, if the threshold is lowered to 0.86, 50% (132) of the primary alcohols are classified correctly. But, unfortunately, 34 compounds not having this group produce a value higher than 0.86 **on this output unit** (false positives).

Any given functional group represents **only a small fraction** of the entire training set; lowering the threshold value would simply generate an excessive number of false positives.

Quite often, such false positives are not considered in the figures of merit calculated for multicategory classifications. Therefore, they have to find a balance between a small percentage of reliable correct predictions and a higher percentage of slightly less reliable ones.

To account for both correct and incorrect assignments, a reliability index called the *A50* value is taken as a measure of the reliability of the prediction for a given functional group *j*:

$$A50 = \frac{0.5\, n_j}{0.5\, n_j + n_{wrong}} \tag{18.3}$$

where n_j is the number of compounds having functional group *j* in the training set, and n_{wrong} is the number of false positives involving group *j*.

NOTE: the *A50* value defines the reliability of the predictions and not the prediction ability. An *A50* value of 100% would mean that at the current threshold level for prediction, only **half** the objects from class *j* are correctly classified, with **no** false positives. This means that if the prediction is positive it is extremely reliable (Figure 18-3), but half of the compounds having this functional group are not identified at all (false negatives).

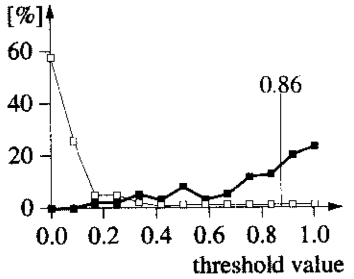


Figure 18-3: Percentage of correct classifications of primary alcohols depending on the threshold value (vertical line); the thin curve shows the percentage of false positives.

At a threshold value of 0.86, 132 of the 265 primary alcohols in the training set are correctly identified, and there are 34 false positives. This gives an *A50* value of 132/(132 + 34) = 79.5%. This is considered a good reliability for predictions. Thirty out of the 36 functional groups are determined with an equally good or better reliability.

These results, obtained with a two-layer neural network (with one hidden layer of 34 neurons), were also compared to an earlier investigation of the same group of authors (Reference 18-2), in which they used only one active layer of neurons, and no hidden layer. It turned out that the hidden layer of neurons leads to remarkable improvements in the reliability of the predictions.

A combination of methods is often more powerful than either one taken separately. Thus, a neural network can be incorporated into an expert system for structure elucidation, where it could help to identify those structures for which functional group identification can be made with a rather high reliability. The search space for the expert system could be narrowed to include only those compounds for which no reliable prediction can be made by the neural network.

A further merit of this work is that even such simple representations of the spectrum and the molecular structure lead to practical results. This opens the door to further investigations aimed at improving the representations of spectral and structural information (See the comment by Bernhard Widrow quoted at the End of Chapter 9.).

18.4 Adjustable Representation of an Infrared Spectrum

A further refinement of the investigation discussed above is to build a number of neural network decision modules and arrange them in a hierarchical manner. In this way, a wide variety of decision possibilities can be achieved, while maintaining a short decision path. Such a work has already been initiated by Kateman and coworkers of the Analytical Department of the Nijmegen Catholic University (see Reference 18-6).

The problem of priorities associated with setting up a hierarchy of decisions was addressed in Section 9.2: how to determine which decision or decisions to put onto the top level of the decision hierarchy, which in the next one, etc. But another, even more important problem is the choice of representation for each of these decision modules.

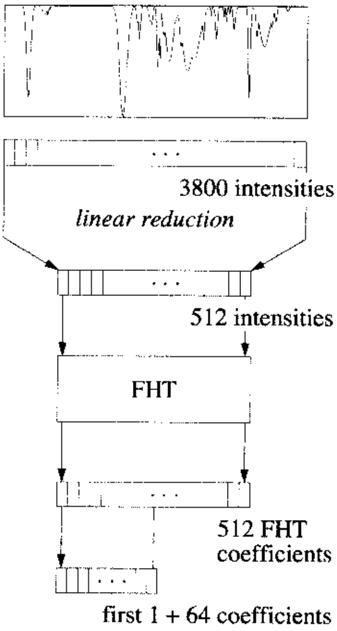


Figure 18-4: The way from the full spectral curve to a reduced set of Hadamard coefficients.

Since the first attempts to build automated interpretation systems for infrared spectra on the basis of the **full spectral curve**, all authors have stressed the fact that different spectral regions are actually used for each decision, so that ideally **a different spectral representation** should be used for each decision. The closest approach to this is offered by a rule-based expert system, which requires a set of rules for each structural feature in the form "**if**-spectral-feature **then** functional-group". Unfortunately, not all the rules necessary to interpret the infrared spectrum have been worked out.

Although most workers are aware of the need for different representations and have stressed it many times, a system that would actually **adapt** spectral input to **each** decision separately has not yet been developed.

In the following two Sections we would like to show a way that possibly enables us to handle both problems: setting the priorities of functional groups (i.e., their positions in the decision hierarchy), and how to select the most suitable spectral representation for a particular decision.

18.5 Representing Spectra using Truncated Sets of Fourier or Hadamard Coefficients

We will discuss how to map spectra from a 512-dimensional spectral space into a two-dimensional structure feature space. We will follow the example of Novic and Zupan of the National Institute of Chemistry in Ljubljana (see Reference 18-3). Their initial representation of an infrared spectrum does not follow the reduction of Equation (18.2); rather, it tries to capture the entire shape of the spectral curve by first making a Hadamard (or Fourier) transformation and then using as input only a **truncated** set (e.g., the first half or quarter) of the coefficients (Figure 18-4).

The Fast Fourier Transformation (FFT) uses as a basis set a set of sines and cosines of different frequencies, while the Fast Hadamard Transformation (FHT) uses box (square wave) functions with different frequencies (Figure 18-5). For the reduction of a measurement space and for recovery of the original information, both transformations have about the same merits and deficiencies (see Reference 18-17).

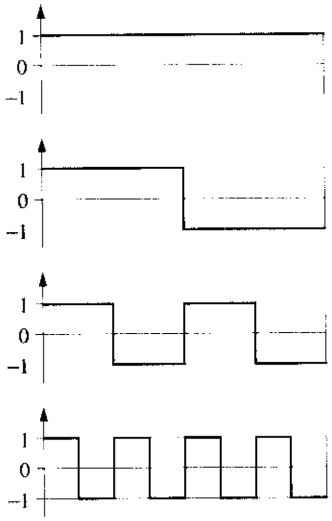


Figure 18-5: A set of square wave (Walsh) functions used in the Hadamard transformation.

Novic and Zupan preferred the Fast Hadamard Transformation because it is 4 to 8 times faster (depending on the hardware) and because it does not use complex coefficients (for a comparison of the algorithms for these two fast transformations see Reference 18-4).

First, the infrared spectrum is divided into 512 intervals, in each of which the corresponding intensity is taken. The intervals are of two different lengths: larger (20 cm^{-1}) at the higher wave numbers (4000 to 2000 cm^{-1}) and narrower (4 cm^{-1}) in the remaining part of the spectrum (2000 to 352 cm^{-1}). Applying the fast Hadamard transformation produces 512 Hadamard coefficients; the first 64 of these are taken as a representation of the spectrum.

This gives us: first, a considerably shorter representation (64 variables compared to 512, a factor of 8), and second, a reasonably good reproduction of the original spectrum. Figure 18-6 shows the same infrared spectrum after its 512 intensities are transformed with the Fast Hadamard Transformation, then reduced, and finally transformed back to the "original".

The **first** of the two goals in this example is to find out which functional groups are so characteristic in the infrared spectra that they can be placed at the top of the decision hierarchy. First, Kohonen maps (Section 7.6) of the 64-variate infrared spectra are made. Then, an attempt is made to associate clusters (of spectra) with common functional groups in the corresponding molecules.

The **second** goal is to obtain guidelines for the **selection of the best spectral representation** for different structures. This task aims to obtain rules (or at least some hints and suggestions) for making new representations appropriate for infrared spectra of compounds having one or more functional groups in common.

As a preliminary example, a modestly large (11 x 11 x 64) Kohonen network (Figure 18-7) was trained with 150 infrared spectra of different compounds (Table 18-1).

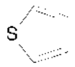
functional group			no. of compounds	label
–O–CO–		*(ester)*	25	E
–CO–		*(ketone)*	21	K
–C–O–C–		*(ether)*	17	O
–COOH		*(acid)*	15	A
		*(thiophene)*	9	T
ketone	+	ether	4	k = K + O
ketone	+	thiophene	3	v = K + T
ketone	+	acid	2	b = K + A
ketone	+	ester	2	e = K + E
ether	+	ester	2	c = O + E
acid	+	ether	1	a = A + O

Table 18-1: Most common functional groups in the training set, and some of the compounds having two functional groups.

The training set contained only compounds having one or two functional groups known to have clearly distinguishable spectral features. Table 18-1 lists some of the functional groups in the compounds whose spectra are input to the Kohonen network.

NOTE: in this application Kohonen maps were obtained **without** toroidal boundary conditions; see Section 6.2 for more details.

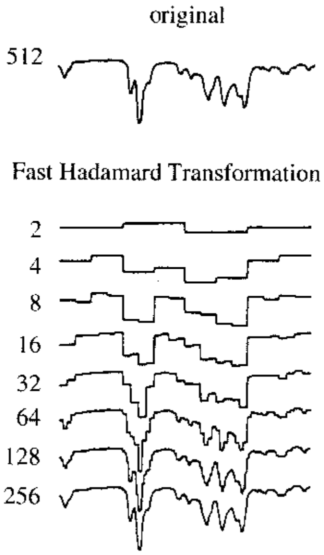


Figure 18-6: Reproduction of the same infrared spectrum from different forms of compression. First the 512 intensities are transformed with the Fast Hadamard Transformation, then reduced, and finally transformed back to the "original".

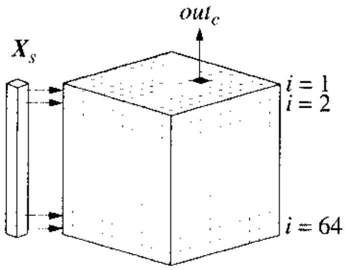


Figure 18-7: Kohonen network for grouping infrared spectra.

18.6 Results of Kohonen Learning

One epoch represents input of all 150 sets of 64 Hadamard coefficients into the (11 x 11 x 64) network. Different numbers of epochs, from 20 to 140, were tried; as a measure of how well the network had adapted, the error produced by all 150 spectra was calculated from all weights of the excited neuron:

$$\varepsilon^{tot} = \frac{\sqrt{\sum_{s=1}^{150}\sum_{i=1}^{64}\left(x_{si}-w_i^{excited}\right)^2}}{\sqrt{150}} \qquad (18.4)$$

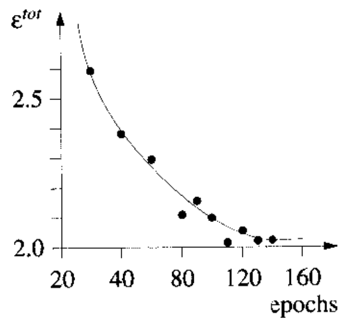


Figure 18-8: Total error ε^{tot} as a function of epochs used in training.

About 100 epochs are needed before the network stabilizes. Figure 18-8 shows a plot of the error at the end of training vs. the number of epochs used. So, no essential improvement can be obtained by going beyond 100 epochs.

In training, the learning rate constant η (Equation (6.6)) was changed linearly, from 0.5 at the beginning to 0.1 at the end, regardless of how many epochs were used.

After the end of training each neuron in the (11 x 11) map was labeled with the letters for functional groups in the compound corresponding to the spectrum that excited it. If **several** spectra excite the same neuron, **all** their labels are attached to its location on the map. The labeled map is shown in Figure 18-9.

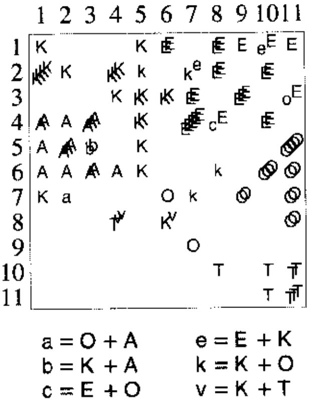


Figure 18-9: The (11 x 11) Kohonen map with some of the 150 labels (each label marks one compound) whose spectrum (via the functional group identified by the label) excited the neuron at that position.

Kohonen learning did produce a map with distinct clusters of some of the labels, which suggests that we can decide how to place the functional groups in a decision hierarchy (of networks) by inspecting these clusters.

The most compact clusters are esters (E) and acids (A); the other two largest sets of compounds, ketones (K) and ethers (O), also form clusters, but have a few outliers; that is, a few spectra labeled "K" or "O" excite neurons outside (but near to) the respective clusters.

Addressing our first goal, the selection of functional groups for the top of a decision hierarchy, E (esters) and A (acids) would seem to be natural choices.

It is interesting to observe the locations of spectra from compounds having **two** functional groups. For example, the two compounds labeled "k" contain both a ketone and an ether, and those labeled "b" contain both the ketone and an acid. These compounds are

> For building the decision hierarchy with **adjusted spectral representations** both types of spectral regions, those of smallest and of largest similarity, are important.

placed between the corresponding clusters, right on the borders where the two clusters of the individual functional groups meet.

As for our second goal, finding guidelines for representing spectra appropriate for individual functional groups, this information is stored in the weights of the 121 neurons.

The 64 weights of each neuron were obtained during training from the truncated set of 64 Hadamard coefficients used to represent the infrared spectrum. Therefore, the "adjusted" weights can be regarded as *"adjusted" Hadamard coefficients* and can consequently be transformed back with the inverse Hadamard transformation to *"adjusted" infrared spectra.*

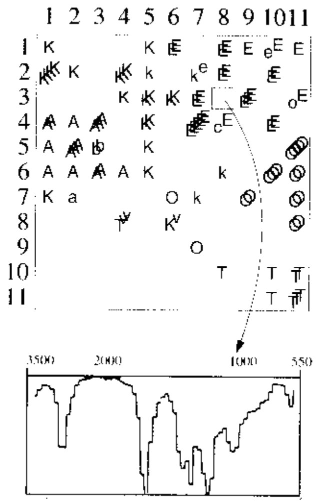


Figure 18-10: Top: Empty space (neuron) which was not excited by any of the training spectra in the "ester" region. Bottom: the corresponding adjusted spectrum stored as a weight vector in this neuron.

(The inverse Hadamard transformation of a **truncated** set of 64 coefficients yields 512 intensities; however, there are only 64 different ones: each intensity is repeated eight times in a row (64 x 8 = 512). Hence, the "adjusted" infrared spectrum has only 64 different intensities over the entire region.)

Note that these 121 adjusted spectra have now replaced the weights of the neurons, i.e., they can be stored in hypercolumns (as neurons) in the Kohonen network. Because the learning procedure adjusts **all weights in all neurons**, the adjusted spectrum can be obtained even **at positions where the neuron was not excited by any of the training spectra**, i.e., from "empty spaces" (see Section 10.4). One such position is marked by a small square in Figure 18-10.

Consider the group of acids (A) in Figure 18-9. The question now is how to obtain the **spectral** features that contribute most to the identification of the **structural** cluster marked with A's.

The spectra of compounds having a specific functional group in common have similar spectral features in certain regions. The hypothesis is that the adjusted spectra forming the region labeled "A" are more similar to each other than they are to the rest of the adjusted spectra in the network, and our goal is to find the regions where the similarity is smallest, and those where it is largest.

The spectral region with a high degree of similarity should be responsible for deciding whether a particular functional group is present, while the parts of the spectrum with lower similarities are responsible for other decisions **afterwards**.

Figure 18-11 shows a comparison between one adjusted spectrum from the left-hand side of the map (acid region) and one from the right-hand side (ether region).

Recall that all 121 adjusted spectra are stored as weights in 64 levels (as hypercolumns) of the (11 x 11) Kohonen network; therefore, it is possible to inspect the **maps of all weight levels**.

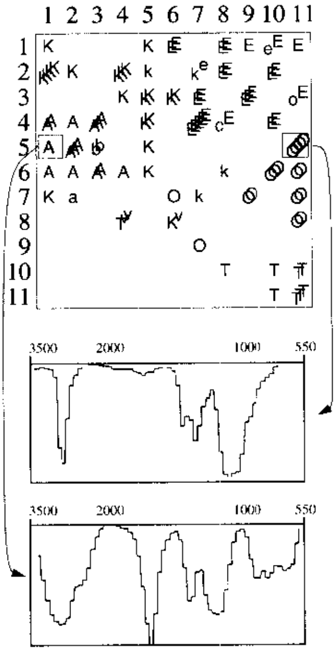


Figure 18-11: Two adjusted spectra from opposite sides of the (11 x 11) Kohonen network.

> Remember that a given level corresponds to a certain wavelength region of the spectrum.

This is done by cutting through the Kohonen network at a certain level (of weights), and plotting a contour map of the intensities at this wavelength region (Figure 18-12 (a)). Figure 18-12)b) shows the map obtained when a cut is made across the network at weight levels 6 and 48.

Taking a closer look at all 64 contour maps, we find that some of the contour lines in certain maps almost exactly coincide with the borders of labeled clusters. Figure 18-13 (b) shows that on level 6 the iso-intensity contours coincide with the cluster labeled A.

Recall that the levels correspond to different spectral regions of "adjusted" spectra. Hence, by finding the levels having intensity contours most similarly distributed to the cluster of a functional group, we will have found the most important spectral regions of this functional group.

If a certain spectral region is highly selective for one particular functional group then, consequently, it is less important to other functional groups. In other words, the **remaining** parts of the spectrum should play a more important role in the next steps, where decisions about **other** functional groups are made.

However, visual comparison of intensity contours using the functional group map is only a preliminary way to find the most relevant spectral regions for a particular decision. A more unbiased way to determine the relevant levels is to compare statistical quantities such as means and standard deviations of weight values (intensities) between the regions inside and outside the cluster.

First, we should determine the area on the map for which a statistical comparison with the outside area has to be made. Figure 18-13 (c) shows the (11 x 11) mask that identifies the "acids" region (ones) as distinguished from the region "outside" the acids (zeros). This mask is applied in all 64 levels of weights, to select spectra for calculation of the mean intensities and their standard deviations, one for each region. Table 18-2 gives the statistical data for intensity distributions within and outside the "acids" region for 18 levels.

According to statistics, levels with large differences between the mean intensities and levels having large standard deviations of intensities outsider the specified cluster are the most significant. That is, first, a large difference between the mean intensity values inside and outside indicates a good possibility for determining the presence or absence of a particular functional group; a larger mean intensity **inside** the cluster than **outside** means that the presence (absence) of a peak is strongly correlated with the presence (absence) of the functional group in the compound. On the other hand, a larger mean intensity **outside** the cluster than **inside** it would mean that the absence of a peak is strongly correlated with the presence of the functional group.

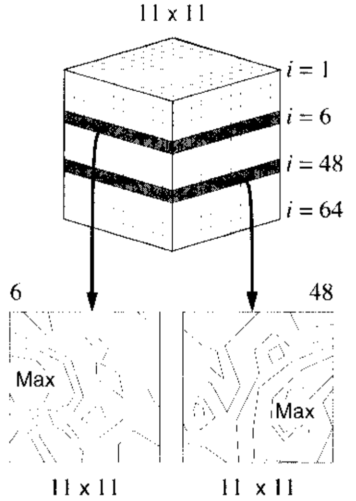


Figure 18-12: Above: if the Kohonen network is cut across different levels, contour plots of intensities can be obtained. Below: these contours are obtained when inspecting levels 6 (b) and 48 (c) corresponding to wave number regions 3190 –3110 and 1100 – 1068 cm^{-1}.

At approximately equal values of mean intensities within and outside the cluster, a **large standard deviation outside** the cluster indicates a spectral region that is rich in information about structural features other than the functional group specified by the cluster.

From Table 18-2 it can be seen that levels 3 to 7 (3310 – 2990 cm^{1}) can be used for recognizing the acids because the mean intensity inside the A region is large compared to the mean intensity outside. On the other hand, the spectral regions represented by levels 26 to 30 (1788 to 1660 cm^{-1}), 36 to 39 (1468 to 1372 cm^{-1}) which have means of intensities and standard deviations of comparable sizes, not too much different to each other, are quite irrelevant for the decision about the acid functional group.

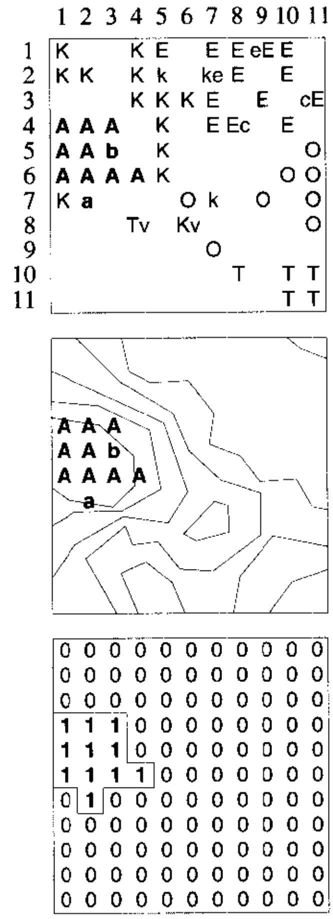
Additional statistical calculations and visual inspection can yield a great deal of information about the correlation between functional groups (represented by clusters in the Kohonen map) and spectral regions (represented by the levels across which the Kohonen network was cut).

With a larger Kohonen map and thousands of spectra of compounds having a larger assortment of structures, such a study can extend the decision sequence to much deeper hierarchical levels.

In this short example, we have seen yet another way to extract information from a trained neural network. The trouble with textbook examples, of course, is that they make it all seem so easy (!).

level	wave no.	w_{av}^{in}	σ_{av}^{in}	w_{av}^{out}	σ_{av}^{out}	Δw_{av}^{in-out}
1	3470	3.22	1.28	1.10	1.19	2.11
2	3390	4.99	1.63	1.35	1.85	3.64
3	3310	6.46	1.58	1.28	1.91	**5.18**
4	3230	7.32	1.22	1.31	1.72	**6.01**
5	3150	7.67	0.93	1.90	1.64	**5.77**
6	3070	7.87	0.74	2.44	1.42	**5.43**
7	2990	8.89	0.57	2.45	1.96	**6.44**
8	2910	7.95	0.67	5.39	2.40	2.56
26	1788	3.73	1.16	1.78	1.52	2.21
27	1756	7.23	0.86	3.86	3.15	3.37
28	1724	9.84	0.51	5.44	3.69	4.40
29	1692	8.13	1.44	4.22	2.70	3.91
30	1660	5.10	1.49	3.37	2.24	1.73
36	1468	4.77	1.34	5.79	1.64	−1.02
37	1276	6.53	0.87	5.97	1.57	0.56
38	1244	6.51	1.13	5.53	1.56	0.98
39	1372	5.22	1.66	5.98	1.75	−0.76
48	1084	2.78	1.27	4.86	2.07	0.71
49	1052	2.15	1.60	4.76	2.10	−2.61
50	1020	1.61	1.23	4.92	2.27	−3.31
51	988	1.79	0.55	3.26	1.22	−1.47
61	668	2.70	0.93	1.68	1.29	1.02
62	636	2.37	0.69	1.50	1.08	0.87
63	604	1.75	0.80	1.85	1.06	−0.10
64	572	0.72	0.85	1.24	0.79	−0.52

Table 18-2: Means of weights, w_{av}, and standard deviations, σ, of intensities inside and outside the masked area (see Figure 18-13 (c)) for some of the levels in the (11 x 11 x 64) Kohonen network of infrared spectra.



	1	2	3	4	5	6	7	8	9	10	11
1	K			K	E		E	E	eE	E	
2	K	K		K	k		ke	E		E	
3				K	K	K	E		E		cE
4	A	A	A		K		E	Ec		E	
5	A	A	b		K						O
6	A	A	A	A	K					O	O
7	K	a				O	k		O		O
8				Tv		Kv					O
9							O				
10								T		T	T
11										T	T

0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

Figurc 18-13: The resulting Kohonen map of functional group labels (a), the map of weights on level six (b) and the mask (c) for sampling the intensities from inside (ones) and outside (zeros) the acids region A (see Figure 18-9).

However, be warned that exquisite care must be taken when applying this procedure for finding distinguishable functional groups and the corresponding spectral regions relevant to adjusting the representation. For each node in the decision hierarchy, a new Kohonen network, new (carefully selected) data, and a new learning procedure are required.

In summary, the described procedure for setting up a decision hierarchy must be followed at **each** decision node. Two steps have to be performed each time: first, on the basis of a Kohonen map, a selection of functional groups for the next decision must be made; second, after the functional groups are selected, the corresponding spectral representations must be determined by inspecting the maps of weights and comparing certain statistical features within and outside the cluster regions.

18.7 A Molecular Transform of the 3D Structure

Clearly, a representation of the structure of a molecule by a set of functional groups, which by necessity is limited, must be insufficient to model the intricate details between structure and infrared spectra. Infrared spectroscopy monitors the vibrations of a molecule in 3D space and a sophisticated modeling of these relationships must take account of the 3D structure of a molecule. As emphasized at several places in this book, the objects of a neural networks study have to be represented by the same number of descriptors (Figure 18-14). This precludes the use of Cartesian coordinates for representing the 3D structure of a molecule because, then, the number of descriptors is directly related to the number of atoms in this molecule. The solution to the problem came by resorting to equations used to derive the 3D structure from an electron diffraction experiment (Figure 18-15).

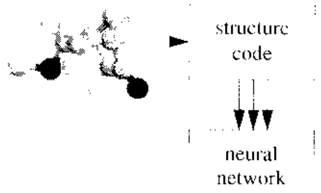


Figure 18-14: Transforming the 3D structure of a molecule into a fixed-length representation.

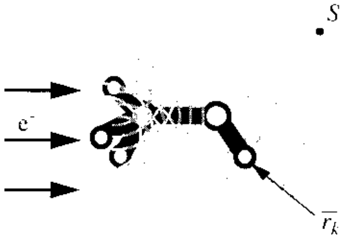
The intensity, $\boldsymbol{I}(s)$, of the electron beam scattered by a molecule at a certain angle, s, is related to the 3D structure of a molecule, represented by the vectors, r_{ij}, of the interatomic distances between atoms i and j, as given by Equation (18.5):

$$\boldsymbol{I}(s) = \sum_{j=i+1}^{n} \sum_{i=1}^{n-1} a_i a_j \frac{\sin(s r_{ij})}{s r_{ij}} \tag{18.5}$$

In this equation, s measures the scattering angle given by

$$s = 4\pi \sin(\theta/2)/\lambda \tag{18.6}$$

with θ being the scattering angle and λ the wave length. a_i and a_j are the atomic numbers of atoms in the molecule. Equation (18.5) is the simplified form of the original equation obtained by assuming the atoms to be point scatterers and the molecule to be rigid. All



S observation point

$\overline{r}_k$ atomic coordinates

Figure 18-15: An electron diffraction experiment.

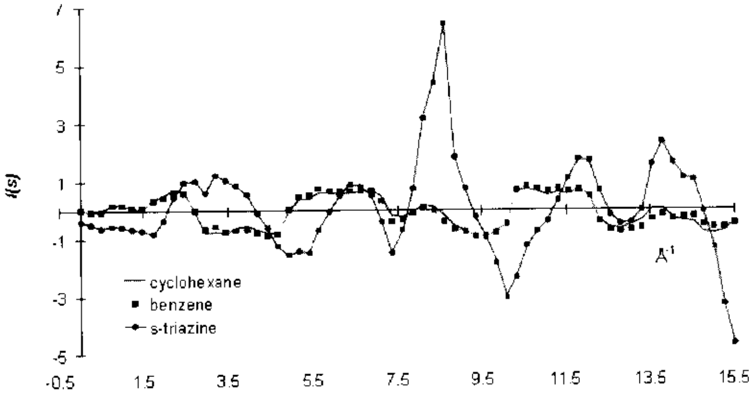


Figure 18-16: The representation of the 3D structure by the 3D-MoRSE code.

instrument variables were collected into a single constant which was set to 1.

In electron diffraction, the scattered intensity, $I(s)$, is measured, and the 3D structure of a molecule, given by r_{ij}, is derived therefrom. We, however, turned Equation (18.5) around by inputting the 3D structure of a molecule and calculating $I(s)$. In addition, $I(s)$ was calculated only for a fixed number (e.g. 32 or 64) of discrete, equidistant values of s.

Thus, the 3D structure was transformed into a fixed number of descriptors. Furthermore, it is possible to use atomic properties, a_i other than atomic numbers in Equation (18.5). In most applications for the simulation of infrared spectra we use partial atomic charges for a_i as calculated by the PEOE method. Figure 18-16 shows such a representation of a molecule. As this molecular representation was derived from an analysis of electron diffraction experiments, we named it 3D-MoRSE code (3D-Molecule Representation of Structure derived from Electron diffraction).

Clearly, such an approach requires access to the 3D coordinates of a molecule. Although X-ray structures have been determined for about 140,000 organic and organometallic compounds, this number is still very small in comparison to the about 14 million known organic compounds. In order to allow the study of the relationships between infrared spectra and the 3D structure of molecules on a broad range, a more universal access to the 3D structure of a molecule is required.

Fortunately, in recent years, automatic 3D structure generators have been developed that can build a 3D model of a molecule from constitutional information embodied in a connection table. One such automatic 3D structure generator is CORINA that has been shown to have a broad scope by automatically converting more than 99% of a database of over 6.5 million structures into 3D coordinates. CORINA can be accessed on the internet

(*http://www2.ccc.uni-erlangen.de/software/corina/*);
see the Appendix for further details.

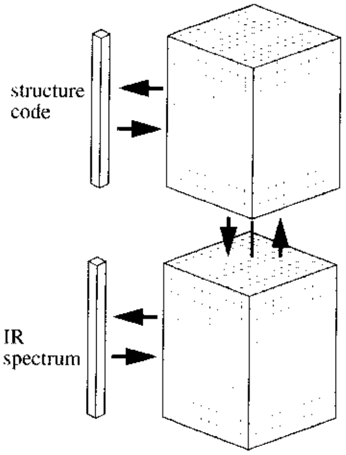


Figure 18-17: Architecture of the counter-propagation network for learning the relationships between 3D structure and infrared spectra.

18.8 Learning by Counter-Propagation

The previous example in this chapter, discussed in Sections 18.5 and 18.6, has shown the merits of a Kohonen network for storing infrared spectra of similar compounds. Similarity of structures, however, was measured rather rudimentary, by a small number of functional groups, only. With the structure coding developed in section 18.7 we have a much more sophisticated structure representation and are now in a position to model the relationships between infrared spectra and structure. We must therefore use a supervised learning technique. In addition, we wanted to retain the advantages of the learning technique embodied in a Kohonen network. The answer is therefore: use a counter-propagation neural network, as this is a supervised learning method using the competitive learning technique also contained in a Kohonen network.

Figure 18-17 shows the architecture of the counter-propagation network used in this study. The upper block of the network contains the weights that are adjusted based on the intensity descriptors, $I(s)$, derived from the 3D structure of the molecules according to Equation (18.5). The lower block contains the weights that are adjusted from the infrared spectra represented as detailed in Section 18.5.

In a typical example, all mono-, di-, and tri-substituted benzene derivatives carrying substituents with no more than eight consecutive bonds and consisting of the atoms C, H, N, O, F, Cl, and Br were retrieved from the SpecInfo© database. This provided a data set of 871 benzene derivatives and their infrared spectra. In order to split this data set into a training and a test set, a planar counter-propagation network of size 30 x 30 was trained with the entire data set. From each of the occupied neurons, that molecule was selected for the training set that had a structure code most similar to the weights of this neuron. This provided a training set of 487 molecules. All other molecules (384) were transferred into the test set. The 3D structure of all molecules was generated by CORINA and converted into 32 intensity coefficients according to Equation (18.5) using partial atomic charges as calculated by the PETRA package as atom descriptors, a_i. The infrared spectrum was reported by 128 intensity values after Hadamard transformation as described in Section 18.5.

A counter-propagation network with toroidal topology consisting of 25 x 25 neurons was trained with these 487 molecules and their infrared spectra.

The similarity between the experimental and the simulated infrared spectra was measured by the correlation coefficient. Figure 18-18 gives the distribution of the correlation coefficients for the entire training set of 487 molecules.

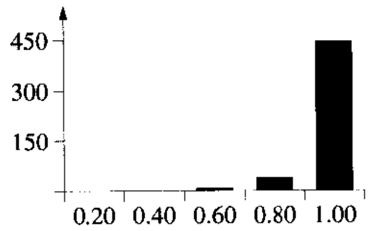


Figure 18-18: Distribution of correlation coefficients of the experimental with the simulated infrared spectra of the 487 molecules of the training set of benzene derivatives.

Clearly, the performance of this approach has to be measured by the test set. Figure 18-19 shows the distribution of the correlation coefficient for the entire test set of 384 compounds.

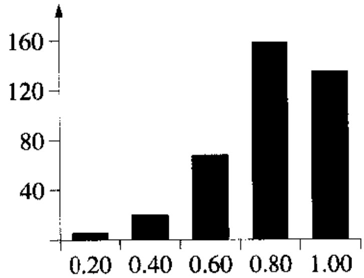


Figure 18-19: Distribution of the correlation coefficients of the experimental with the simulated infrared spectra of the 384 molecules of the test set of benzene derivatives.

Figure 18-20 compares one of the higher quality simulated infrared spectrum with the experimental one from the test set. The most important result is that good correspondence can be obtained over the entire frequency range, not only in the region of valence bond vibrations but also in the fingerprint region. This attests to the potential of the 3D-MoRSE code to represent the entire structure of a molecule in its coefficients, not only parts of the structure such as functional groups.

As with any learning procedure, the quality of prediction is dependent on the availability of information. We have found that the cases of predictions of poorer quality can, by and large, be attributed to a lack of data for these types of compounds. However, all cases with correlation coefficients of 0.7 and higher can be considered for many applications as satisfactory.

18.9 Different Strategies for the Selection of a Training Set

In the example given in Section 18.8, a certain group of compounds, in this case, mono-, di-, and tri-substituted benzene derivatives, was selected as data set to train a counter-propagation network for the simulation of infrared spectra.

It is also possible to train a single counter-propagation network for the simulation of infrared spectra over the entire range of organic

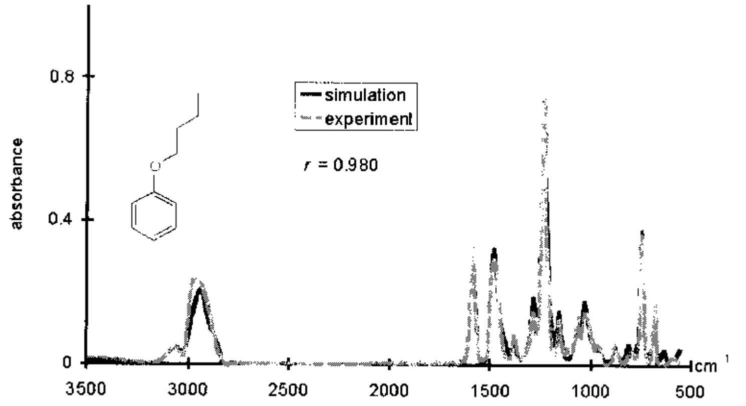


Figure 18-20: Comparison of a simulated with an experimental infrared spectrum for a molecule of the test set of benzene derivatives.

chemistry. The SpecInfo database Version 2.0 contained approximately 15,000 infrared spectra. After elimination of duplicates and ionic species we were left with 9,850 different structures. This data set was split into 3,244 compounds for training a counter-propagation network with 70 x 70 neurons, and two test sets of about equal size.

By and large the infrared spectra simulated with this large comprehensive neural network were quite acceptable. Clearly, however, the computation times for training such a huge network were quite high and the results with smaller, dedicated networks were often of higher quality. Thus, we rather prefer smaller networks.

In both approaches, working with a large network encompassing the entire range of organic chemistry or when training a network for a certain class of compounds, the training of the network can be done once and for all; predictions are then rather rapid, indeed.

However, it must be realized that the selection of a compound into a class of compounds might be quite arbitrary. Is the compound shown in Figure 18-21 a quinoline derivative or a substituted furane? (Clearly, it is both!) For such cases, we have developed an alternative for the selection of a data set for training a network: The structure for which an infrared spectrum should be simulated, the query structure, determines its own training set: From the database those 50 molecules – and their associated infrared spectra – are selected for the training of a counter-propagation network, that have a structure code that is most similar to the code of the query structure. We have found this approach to give the most satisfactory results although in many cases,

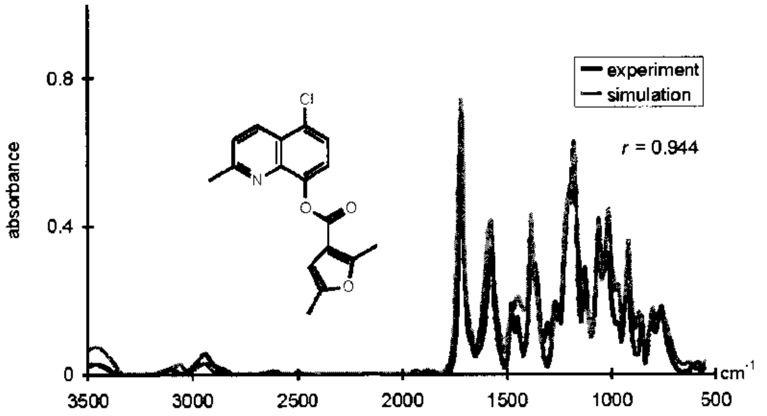


Figure 18-21: Comparison of an experimental infrared spectrum with the simulated spectrum obtained by the query-directed approach from a counter-propagation network trained with the 50 most similar structures.

where the definition of a class of compounds is quite clear, the improvements are only minor.

Figure 18-21 gives the results of such a query-directed simulation of an infrared spectrum. The disadvantage of this approach is that each new query structure requires the training of a specific network, one cannot work with pre-trained networks. However, with training sets of 50 compounds the training times are quite acceptable with about a minute on a PC.

18.10 From the Infrared Spectrum to the 3D Structure

The architecture of Figure 18-17 shows that there is a direct relationship between the block of weights obtained from the structure code and the block of weights from the infrared spectra. As of now, we have used such a counter-propagation network in a single direction, inputting a structure code and outputting an infrared spectrum.

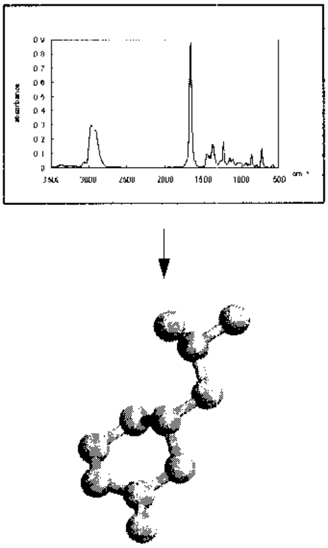


Figure 18-22: 3D structure derived from an infrared spectrum.

However, it should be possible to also operate such a counter-propagation network in reverse mode, inputting an infrared spectrum and outputting a structure code (Figure 18-22). Now, remember, that the structure code as calculated by Equation (18.5) is nothing else than a discrete form of the electron diffraction pattern, the very information that is used – in its full form – to derive a 3D structure from an electron diffraction experiment. Thus, it should be possible to transform the structure code obtained from the counter-propagation network into a 3D structure. In fact, we have succeeded to develop such a method based on radial distribution functions, a structure code quite similar to the 3D-MoRSE code (see Section 21.3). Figure 18-22 shows an example of a 3D structure directly predicted from the infrared spectrum. This is the first time ever that is has become possible to derive a 3D structure from an infrared spectrum. A discussion of this procedure goes beyond the scope of this book. The interested reader is referred to the original publications in journals.

The important message to carry away is that novel information can be gained with a sophisticated structure code and a powerful learning algorithm such as the one embodied in a counter-propagation network.

The methods explained in Sections 18.7 – 18.10, both for the simulation of infrared spectra and for the derivation of the 3D structure from an infrared spectrum, are made available for the general public to use on the internet through the project TeleSpec at

http://www2.ccc.uni-erlangen.de/research/ir/.

Furthermore, the scientific community is invited to use and cooperate in building a freely accessible database of infrared spectra. See the Appendix for further details.

18.11 References and Suggested Readings

18-1. M. E. Munk, M. S. Madison and E. W. Robb, "Neural Network Models for Infrared Spectrum Interpretation", *Mikrochim. Acta* [Wien] 1991 II, 505 – 514.

18-2. E. W. Robb and M. E. Munk, "A Neural Network Approach to Infrared Spectrum Interpretation", *Mikrochim. Acta*, [Wien], 1990 I, 131 – 155.

18-3. M. Novic and J. Zupan, "2-D Mapping of Infrared Spectra Using Kohonen Neural Network", *Vestn. Slov. Kem. Drus.* **39** (1992) 195 – 212.

18-4. M. Razinger and M. Novic, "Reduction of the Information Space for Data Collection", in *PC's for Chemists*, Ed.: J. Zupan, Elsevier, Amsterdam, NL, 1990, pp. 89 – 103.

18-5. J. R. M. Smiths, P. Schoenmakers, A. Stehmans, F. Sijstermans and G. Kateman, "Interpretation of Infrared Spectra with Modular Neural Network Systems", *Chemom. Intell. Lab. Syst.* **18** (1993) 27 – 39.

18-6. W. J. Melssen, J. R. M. Smits, G. H. Rolf and G. Kateman, "Two-dimensional Mapping of Infrared Spectra Using Parallel Implemented Self-organising Feature Map", *Chemom. Intell. Lab. Syst.* **18** (1993) 195 – 204.

18-7. J. Zupan and M. Novic, "Hierarchical Ordering of Spectral Data", in *Computer Supported Spectroscopic Data Bases*, Ed.: J. Zupan, Ellis Horwood, Chichester, UK, 1986, pp. 42 – 63.

18-8. M. Tusar and J. Zupan, "Neural Networks", in *Software Development in Chemistry 4*, Ed.: J. Gasteiger, Springer Verlag, Berlin, FRG, 1990, pp. 363 – 376.

18-9. M. Otto and U. Hörchner, "Application of Fuzzy Neural Networks to Spectrum Identification", in *Software Development in Chemistry 4*, Ed.: J. Gasteiger, Springer Verlag, Berlin, FRG, 1990, pp. 377 – 384.

18-10. J. R. Long, V. G. Gregoriou and P. J. Gemperline, "Spectroscopic Calibration and Quantization Using Artificial Neural Networks", *Anal. Chem.* **62** (1990) 1791 – 1797.

18-11. B. Curry and D. E. Rumelhart, "MSnet: A Neural Network That Classifies Mass Spectra", *Tetrahedron Comput. Methodol.* **3** (1990) 213 – 237.

18-12. J. U. Thomsen and B. Mayer, "Pattern Recognition of the ^{1}H NMR Spectra of Sugar Alditols Using a Neural Network", *J. Magn. Res.* **84** (1989) 212 – 217.

18-13. H. Lohninger, "Classification of Mass Spectral Data Using Neural Networks", in *Software Development in Chemistry* **5**, Ed.: J. Gmehling, Springer Verlag, Berlin, FRG, 1991, pp. 159 – 168.

18-14. B. J. Withoff, S. P. Levine and S. A. Sterling, "Spectral Peak Verification and Recognition Using a Multilayered Neural Network", *Anal. Chem.* **62** (1990) 2709 – 2719.

18-15. P. J. Gemperline, J. R. Long and V. G. Gregoriou, "Nonlinear Multivariate Calibration Using Principal Component Regression and Artificial Neural Networks", *Anal. Chem.* **63** (1991) 2314 – 2323.

18-16. V. Kvasnicka, "An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts", *J. Math. Chem.* **6** (1991) 63 – 76.

18-17. J. Zupan, *Algorithms for Chemists*, Wiley, Chichester, UK, 1989, Chapter 5.

18-18. J. Gasteiger, X. Li, V. Simon, M. Novic and J. Zupan, "Neural Nets for Mass and Vibrational Spectra", *J. Mol. Struct.* **292** (1993) 141 – 160.

18-19. J. Sadowski and J. Gasteiger, "From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders", *Chem. Reviews* **93** (1993) 2567 – 2581.

18-20. J. Sadowski, J. Gasteiger and G. Klebe, "Model Builders Using 639 X-Ray Structures", *J. Chem. Inf. Comput. Sci.* **34** (1994) 1000 – 1008.

18-21. M. Novic and J. Zupan, "Investigations of Infrared-Spectra-Structure Correlation Using a Kohonen and Counterpropagation Neural Networks", *J. Chem. Inf. Comput. Sci.* **35** (1995) 454 – 466.

18-22. J. H. Schuur, P. Selzer and J. Gasteiger, "The Coding of the Three-dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure - Spectra Correlations and Studies of Biological Activity", *J. Chem. Inf. Comput. Sci.* **36** (1996) 334 – 344.

18-23. J. Schuur and J. Gasteiger, "Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a Novel 3D Structure Representation", *Anal. Chem.* **69** (1997) 2398 – 2405.

18-24. The SpecInfo© database was provided by Chemical Concepts GmbH, Weinheim, Germany.

18-25. M. C. Hemmer, V. Steinhauer and J. Gasteiger , "The Prediction of the 3D Structure of Organic Molecules from Their Infrared Spectra", *Vibrat. Spectroscopy* **19** (1999) 151 – 164.