

Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns Using Neural Networks

U.-M. Weigel and R. Herges*

Institute of Organic Chemistry, University of Erlangen-Nürnberg, Henkestrasse 42, 8520 Erlangen, Germany

Received July 14, 1992

Several neural networks of back-propagation type were trained to recognize and to identify the substitution patterns of differently substituted aromatic compounds by pattern analysis of their infrared spectra. The effect of network architecture and input design on the predictive performance of the neural net was examined. The neural nets were trained with a set of IR spectra of mono-, di (1,2; 1,3; 1,4)-, tri (1,2,3; 1,3,5; 1,2,4)-, and tetrasubstituted (1,2,3,4; 1,2,3,5; 1,2,4,5) aromatic compounds. After training they were able to interpret spectra of aromatic compounds not included in the training set. The accuracy in recognizing the proper substitution pattern ranges from fair to excellent, depending on the model and the training set.

INTRODUCTION

Starting with the development of the first artificial neural networks in 1943 by McCulloch and Pitts,¹⁹ the research in this field was at first restricted to a limited group of mathematicians and computer scientists. The situation changed with the introduction of the Hopfield²⁰⁻²² model (1982) and the back-propagation^{23,11} algorithm (1986). This boosted the number of applications in science including chemistry.

Neural networks proved to be useful in the following main areas: pattern recognition, classification, and parametrization. Applications of neural networks in chemistry focus on models of knowledge representation in chemical engineering,¹ protein secondary structure analysis and prediction,² DNA promoter site recognition,³ prediction of electrophilic aromatic substitution,⁴ and mainly spectroscopy (¹H NMR,⁵ IR,^{6,7} UV, MS). Zupan and Gasteiger⁸ gave an excellent review concerning both neural network models and their applications.

Two articles that appeared recently,^{6,7} describe the recognition of functional groups by pattern recognition from IR spectra, using the back-propagation algorithm. The characteristic peaks of functional groups in the IR spectrum are routinely used for structure elucidation in analytical organic chemistry. Both papers describe simple neural nets, without⁶ or with one⁷ hidden layer, that were trained to recognize most of the common functional groups from these peculiar absorption bands with high accuracy. More subtle structural features, however, are more difficult to interpret (by man as well as computers), and probably more sophisticated network models are needed to gain structural information beyond functional group recognition. In favorable cases, the substitution pattern of aromatic compounds can be deduced from characteristic absorption patterns in the IR.¹² However, interpretation becomes difficult if the spectra are not measured at a very high concentration or if the characteristic absorption bands are altered or obscured by other structural features. Characteristic bands in the fingerprint region (900–600 cm⁻¹) are often absent or overlap with patterns of differently substituted aromatics. Thus, interpretation is intricate and very often ambiguous. On the other hand, there exists an extensive and consistent set of spectra of substituted aromatic compounds that can be used for the training of artificial neural networks. Since neural networks learn and derive rules for

Table I. Characteristic Fingerprint Absorption Bands¹²

| substitution type | range of absorption in region II (cm ⁻¹) | | |
|-------------------|--|----------------------|----------------------|
| mono | 690–710 ^a | 730–770 ^b | |
| di | | | |
| [1,2] | | 735–770 ^b | |
| [1,3] | 680–725 ^a | 750–810 ^b | 810–900 ^b |
| [1,4] | | | 800–860 ^b |
| tri | | | |
| [1,2,3] | 660–720 ^a | 750–810 ^b | |
| [1,3,5] | 660–735 ^a | 800–850 ^b | 830–875 ^b |
| [1,2,4] | 660–720 ^a | 800–860 ^b | 810–900 ^b |
| tetra | | | |
| [1,2,3,4] | | | 800–900 ^b |
| [1,2,3,5] | | | 800–900 ^b |
| [1,2,4,5] | | | 800–900 ^b |

^a CC wagging vibration. ^b CH wagging vibration.

Table II. Input Types A, B, and C

| type | region I (2000–1400 cm ⁻¹) | | region II (900–600 cm ⁻¹) | | total no. of input neurons |
|------|--|----------------|---------------------------------------|----------------|----------------------------|
| | interval (cm ⁻¹) | N ^a | interval (cm ⁻¹) | N ^a | |
| A | 12.5 | 48 | | | 48 |
| B | 12.5 | 48 | 12.5 | 24 | 72 |
| C | 12.5 | 48 | 6.25 | 48 | 96 |

^a N = number of input neurons.

examples, this seemed to be an ideal field of application.

IR SPECTRA OF AROMATIC COMPOUNDS

The substituted benzene skeleton shows characteristic absorption patterns in two regions of the infrared spectrum¹² dependent on the substitution type:

- region I (a) overtone/combination vibrations (CH vibrations)
2000–1600 cm⁻¹
(b) C=C stretch vibrations
1600–1400 cm⁻¹
region II fingerprint (CH and CC wagging vibrations)
900–600 cm⁻¹

IR peaks of region Ia result from nonnatural vibrations with a change in quantum number $\Delta\nu > 1$ (overtone

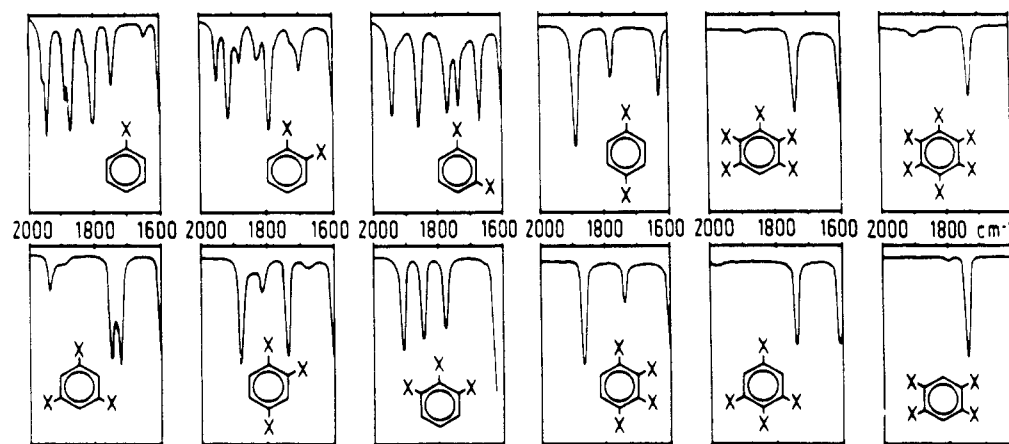


Figure 1. Overtone and combination vibrations (2000–1600 cm^{-1} , CH vibrations) of aromatic substitution types.¹²

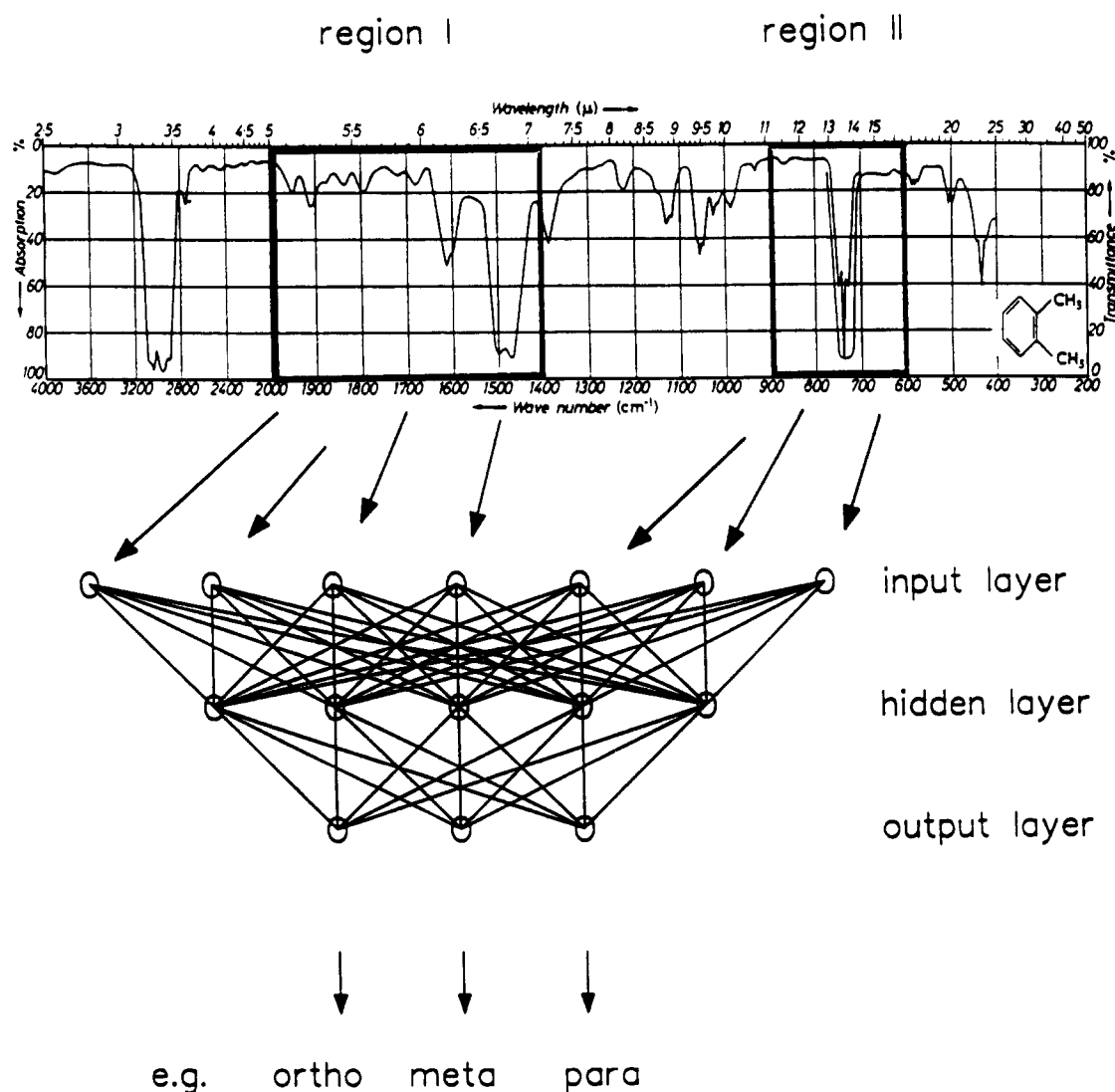


Figure 2. Input design and network architecture.

vibrations: $\Delta\nu = \pm 2, \pm 3$) and combination vibrations. The probability for transition to higher energy levels is small compared to natural vibrations ($\Delta\nu = 1$); therefore, overtone vibrations in general have weak intensities, and samples have to be measured at a high concentration or a thick probe layer. Most spectra of aromatic compounds in spectra databases or reference handbooks, however, show an empty or incomplete region I and, thus, give little or no information. In Figure 1 the characteristic patterns of overtone/combination vibrations for the 10 aromatic substitution types treated here are shown.

C=C stretching vibrations of region Ib (1600–1400 cm^{-1}) usually appear as a group of four bands. They are highly characteristic of the aromatic ring itself but are of limited value for assigning the substitution pattern. In contrast to region I, region II contains absorption bands corresponding to natural vibrations ($\Delta\nu = \pm 1$, CH and CC wagging vibrations) with large intensities, thus usually providing more information to determine the substitution type. Table I shows the characteristic fingerprint absorption bands for the 10 aromatic substitution types.

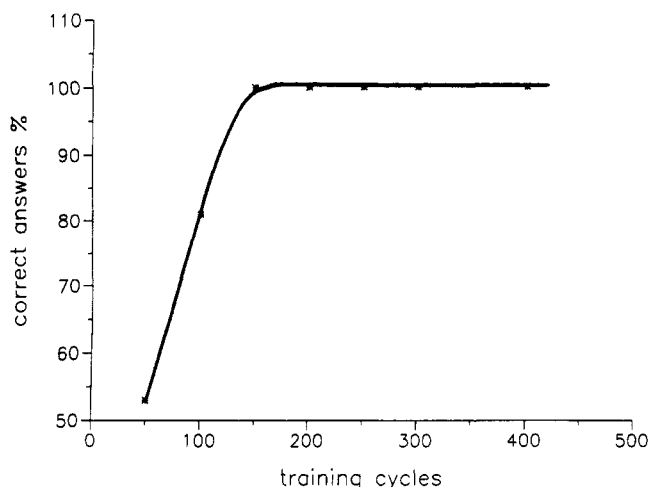


Figure 3. Neural network behavior in the training phase (disubstitution). Correct answers are in percent as a function of the number of training cycles. After n cycles, the complete training set is learned correctly (100%), and network stability is achieved.

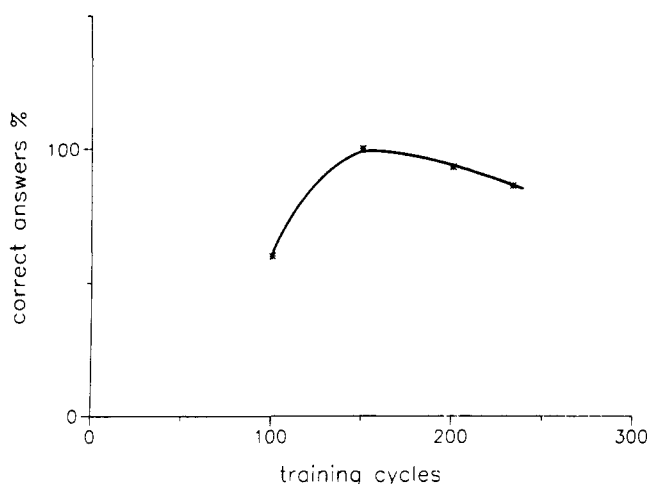


Figure 4. Neural network behavior in the test place (disubstitution), correct answers in percent as a function of the number of training cycles. At first glance, it seems to be obvious that the net should perform best on interpretation of unknown test spectra, if it is thoroughly trained in the training stage. However, the optimum of performance is reached in "incompletely" trained nets. This effect is referred to as "overtraining" and is due to the fact that the capability of generalizing is lost if a training set is perfectly learned.

METHODS

Choice of Network Model. There are two general types of neural nets: autoassociative (unsupervised learning) and heteroassociative (supervised learning). The determination of the substitution type of aromatic compounds from IR spectra is a pattern recognition and classification type problem that in principle can be solved by both types. In a preliminary study we, therefore, chose three different network types: ART2,⁹ Kohonen,¹⁰ and back-propagation.¹¹ The first two are autoassociative; the last one is heteroassociative.

ART2 and Kohonen nets are classifiers. Applied to our problem, the net should classify a number of differently substituted aromatic compounds and attribute these to n self-defined substitution classes. From preliminary tests, however, we learned that this concept is successful only if the spectra within each class show little difference in their characteristic peaks and intensities. In our case spectra of the same class differ extremely in wavelength and intensity. Moreover, characteristic peaks are frequently very weak or absent. This induces the network to create new classes or misleads

Table III. Number of Training and Test Spectra for Experiments within Different Substitution Classes

| substitution classes | pattern no. of training spectra | no. of test spectra | |
|----------------------|---------------------------------|---------------------|----------|
| | | <i>a</i> | <i>b</i> |
| mono | | | 7 |
| di | [1,2] | 10 | 6 |
| | [1,3] | 10 | 4 |
| | [1,4] | 10 | 4 |
| tri | [1,2,3] | 10 | 4 |
| | [1,3,5] | 10 | 3 |
| | [1,2,4] | 10 | 8 |
| tetra | [1,2,3,4] | 10 | 3 |
| | [1,2,3,5] | 10 | 20 |
| | [1,2,4,5] | 10 | 8 |
| mono/di/tri/tetra | 100 | | 78 |

^a Test with each class separately. ^b Experiments with all 10 substitution types.

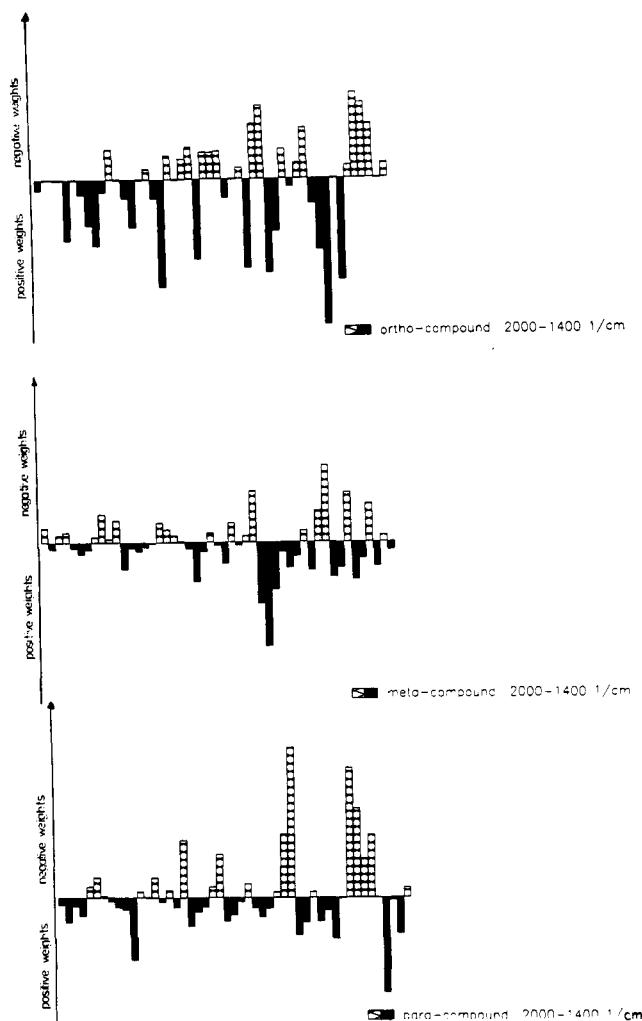


Figure 5. Plot of the weight vector (sum of the contribution of each input unit to a particular output unit) for region I (overtone/combination vibrations) of disubstituted compounds (ortho, meta, para). Solid bars represent positive weights, and hatched bars represent negative weights. The weights can be interpreted in terms of the "learned patterns" for identification of the substitution type.

classification. The heteroassociative back-propagation algorithm proved to be superior for tackling our problem and was used throughout.

Input and Output Design. Since IR spectra recorded on a nondigital spectrometer provide merely analog data, the spectra have to be converted into a computer-readable form for network input. For that purpose we created three types of digitally coded spectra¹⁵ A-C which provide information

Table IV. Prediction Results Obtained with Network Type: $(N^a) \rightarrow [(7) \rightarrow (21)] \rightarrow (3)^b$

| input type | N^a | disubstitution | | | | trisubstitution | | | total |
|------------|-------|----------------|-------|-------|-------|-----------------|---------|---------|-------|
| | | [1,2] | [1,3] | [1,4] | total | [1,2,3] | [1,3,5] | [1,2,4] | |
| A | 48 | 66.7 | 50.0 | 25.0 | 50.0 | 75.0 | 33.0 | 100.0 | 80.0 |
| B | 72 | 50.0 | 50.0 | 100.0 | 64.3 | 75.0 | 100.0 | 62.5 | 80.0 |
| C | 96 | 100.0 | 100.0 | 100.0 | 100.0 | 75.0 | 66.7 | 87.5 | 80.0 |

^a N = number of input neurons. ^b Proportion of correct answers is in percent.

Table V. Disubstitution: Variation of the Number of Hidden Layers^a

| hidden layer | network architecture | [1,2] | [1,3] | [1,4] | total | training cycles ^b |
|--------------|---|-------|-------|-------|-------|------------------------------|
| 0 | (96) \rightarrow (3) | 66.7 | 100.0 | 100.0 | 93.0 | 450 |
| 1 | (96) \rightarrow [(27)] \rightarrow (3) | 50.0 | 100.0 | 100.0 | 85.7 | 241 |
| 2 | (96) \rightarrow [(7) \rightarrow (21)] \rightarrow (3) | 100.0 | 100.0 | 100.0 | 100.0 | 150 |

^a Proportion of correct answers is in percent. ^b One training cycle = one presentation of 30 training spectra to the network.

Table VI. Trisubstitution: Variation of the Number of Hidden Layers^a

| hidden layer | network architecture | [1,2,3] | [1,3,5] | [1,2,4] | total | training cycles ^b |
|--------------|--|---------|---------|---------|-------|------------------------------|
| 0 | (96) \rightarrow (3) | 100.0 | 66.7 | 87.5 | 87.0 | 924 |
| 1 | (96) \rightarrow [(27)] \rightarrow (3) | 75.5 | 66.7 | 87.5 | 80.0 | 198 |
| 2 | (96) \rightarrow [(31) \rightarrow (11)] \rightarrow (3) | 100.0 | 66.7 | 87.5 | 87.0 | 275 |

^a Proportion of correct answers is in percent. ^b One training cycle = one presentation of 30 training spectra to the network.

Table VII. Tetrasubstitution: Variation of the Number of Hidden Layers^a

| hidden layer | network architecture | [1,2,3,4] | [1,2,3,5] | [1,2,4,5] | total | training cycles ^b |
|--------------|---|-----------|-----------|-----------|-------|------------------------------|
| 0 | (96) \rightarrow (3) | 66.7 | 65.0 | 50.0 | 61.3 | 300 |
| 1 | (96) \rightarrow [(27)] \rightarrow (3) | 66.7 | 65.0 | 50.0 | 61.3 | 250 |
| 2 | (96) \rightarrow [(7) \rightarrow (21)] \rightarrow (3) | 66.7 | 60.0 | 63.0 | 61.3 | 300 |

^a Proportion of correct answers is in percent. ^b One training cycle = one presentation of 30 training spectra to the network.

from rough to fine. The infrared regions I (2000–1600 cm^{-1} CH vibrations and 1600–1400 cm^{-1} CC vibrations) and II (900–600 cm^{-1} CH and CC wagging vibrations) were divided into intervals with an increment of 12.5 or 6.25 cm^{-1} ; each of which was assigned to one input neuron. If an absorption band appears within the frequency limits of an interval, the corresponding intensity was assigned to it, otherwise 0 absorption was assumed (Figure 2). An alternative way of coding IR spectra for input used in many databases is peak lists, consisting of peak position and intensity ($\text{cm}^{-1}/\%$). When used as input, the network performance, however, was considerably reduced. Peak lists, therefore, were not used in further experiments. Table II shows the structure of the input types A, B, and C which have been used in this work. Our IR spectra were from the DMS-spectral cards,¹⁵ with 23 000 spectra of organic compounds. The database contains 2061 mono-, 1899 di- (305 [1,2]-, 243 [1,3]-, and 1351 [1,4]-substituted aromatics), 431 tri- (83 [1,2,3]-, 48 [1,3,5]-, and 300 [1,2,4]-substituted aromatics), 171 tetra-, and 79 penta- and hexasubstituted aromatic compounds (see Table X for a complete list and the *The Aldrich Library of Infrared Spectra*³⁰).

Because of strong overlap of the carbonyl, nitro, and amino groups with CH wagging vibrations, these compounds were

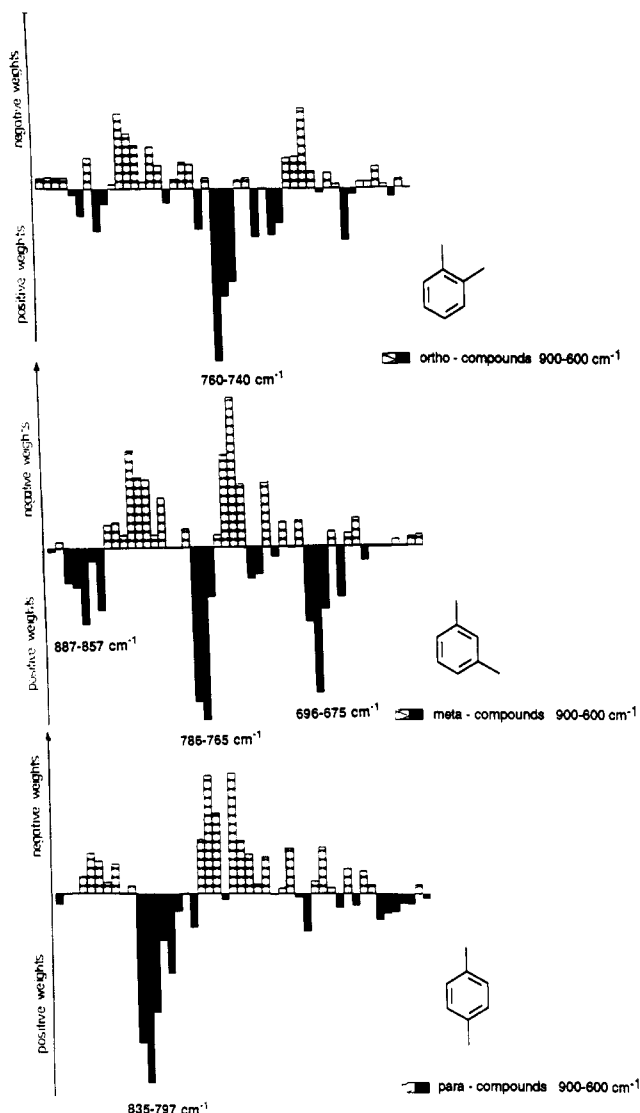


Figure 6. Plot of the weight vector of region II (fingerprint region) for disubstituted compounds. Solid bars represent positive weights, and hatched bars represent negative weights.

excluded from the training set. Spectra are measured as Nujol mulls and KBr pellets.

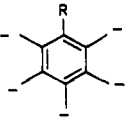
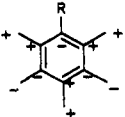
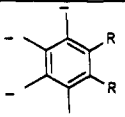
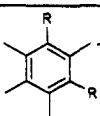
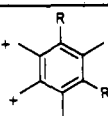
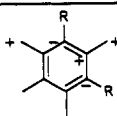
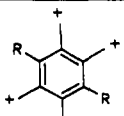
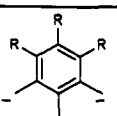
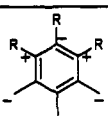
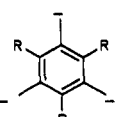
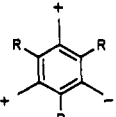
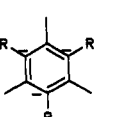
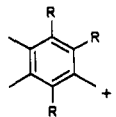
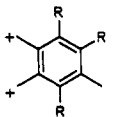
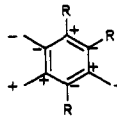
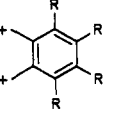
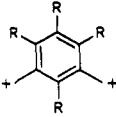
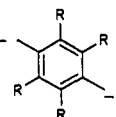
Training and Test Stage. When experimenting with the classes of di-, tri-, and tetrasubstituted aromatic compounds separately (three patterns must be distinguished in each case), the training files always consisted of 30 training spectra, each substitution type contributing 10 spectra of different quality. The network was fed with these spectra in an alternating sequence, e.g., for recognition of disubstituted aromatics, [1,2], [1,3], [1,4], [1,2], [1,3], Table III gives detailed information about training and test files.

The smallest group of available spectra are the [1,2,3,4]-tetrasubstituted aromatics with 13 examples. Since the number of training spectra for each group should be equal for consistent training, we chose stochastically 10 spectra from each other substitution group for training and 3–6 test spectra.

Figure 2 depicts the training procedure for our back-propagation network. The spectra and the substitution type of the compounds were used as input, and the weights were adjusted according to the back-propagation procedure until a stable set of weights was obtained.

Network Architecture. The number of neurons and the topology of the connections between them represent the network architecture. Since the information contained in a

Table VIII. Characteristic H Out-of-Plane Wagging Vibrations of Differently Chlorosubstituted Aromatics, According to Semiempirical Quantum Chemical Calculations (MNDO)^a

| | | | | | |
|-----------|---|---|---|--|--|
| mono |  |  | | | |
| | $\nu=825.9$ 1/cm T=0.49 | $\nu=825.4$ 1/cm T=0.27 | | | |
| [1,2] |  | | | | |
| | $\nu=839.7$ 1/cm T=0.48 | | | | |
| [1,3] |  |  |  | | |
| | $\nu=991.6$ 1/cm T=0.25 | $\nu=871.6$ 1/cm T=0.39 | $\nu=626.4$ 1/cm T=0.38 | | |
| [1,4] |  | | | | |
| | $\nu=914.7$ 1/cm T=0.54 | | | | |
| [1,2,3] |  |  | | | |
| | $\nu=873.6$ 1/cm T=0.43 | $\nu=741.2$ 1/cm T=0.35 | | | |
| [1,3,5] |  |  |  | | |
| | $\nu=998.2$ 1/cm T=0.46 | $\nu=992.6$ 1/cm T=0.04 | $\nu=623.1$ 1/cm T=0.46 | | |
| [1,2,4] |  |  |  | | |
| | $\nu=991.4$ 1/cm T=0.27 | $\nu=914.4$ 1/cm T=0.4 | $\nu=707.0$ 1/cm T=0.1 | | |
| [1,2,3,4] |  | | | | |
| | $\nu=913.9$ 1/cm T=0.4 | | | | |
| [1,2,3,5] |  | | | | |
| | $\nu=996.2$ 1/cm T=0.4 | | | | |
| [1,2,4,5] |  | | | | |
| | $\nu=986.7$ 1/cm T=0.43 | | | | |

^a Transition dipole (quantum chemical equivalent for absorption) and frequency correspond closely to the characteristic absorptions (weight vectors) found by the neural net (Figures 6–8). R = Cl; ν = vibrational frequency; T = transition dipole.

neural network is stored entirely in the strengths of the connections between the neurons, the choice of the appropriate network architecture is of utmost importance for its efficiency. Figure 2 depicts the general arrangement of the feed forward type neural net,¹¹ which was used in the present study. The number of hidden layers (0, 1, 2) and the number of the neurons in each layer were varied to investigate the implications on network efficiency.

Throughout this paper network architecture is defined as follows:

(input neurons) \rightarrow [(hidden layer 1 neurons) \rightarrow (hidden layer 2 neurons) \rightarrow ...] \rightarrow (output neurons)

Network Parameters. All networks in the present study were trained with the following parameter set, using the commercial neural network simulator ANSIM.¹³ The training and test files were normalized between -0.5 and 0.5 according to the software recommendations.

noise: 0.7

decay of noise with every pass through input file: 0.01

learning rate: 0.1

Network training was finished when the following data were reached:

RMS error: 0.01

max output unit error: 0.01

Thoroughly trained networks (100% correct answers in the training step, Figure 3) do not perform at their optimum level.

In this stage the network has learned the input spectra *exactly*. Generalization ability and, therefore, prediction performance on unknown spectra are reduced (Figure 4). For tests with unknown spectra, we used trained networks with the following data:

RMS error: 0.1

max output unit error: 0.2

Using these parameters we reached the optimum level for prediction in the test stage (Figure 4).

RESULTS

Effect of Input Design on Prediction Performance. The selection of the appropriate spectral region containing the essential structural information and the digitizing procedure for preparing the input vector are the most important variables in input design. Table IV shows the performance of the net upon variation of the input type (for definition of input types A–C, see Table II).

Input type A, in which the fingerprint region II (900–600 cm^{-1}) is omitted, gave poor results in predicting the substitution patterns of disubstituted aromatics (50% correct answers, Table IV). This is probably due to the fact that the IR spectral region I (2000–1400 cm^{-1}) usually is not perfectly resolved or recorded at a sufficiently high concentration and per se does not provide the necessary information for unambiguous interpretation. Additional information from region II (900–

Table IX. List of Spectra Used for Training and testing

| no. ^a | compd | no. | compd |
|-----------------------|---|------------------------|--|
| 3340 | iodobenzene (v) | 3338 | chlorobenzene (v) |
| 3344 | phenol (v) | 3339 | bromobenzene (v) |
| 4837 | phenol (l) | 218 | butylbenzene (l) |
| 221 | isobutylbenzene (l) | 220 | <i>tert</i> -butylbenzene (l) |
| 572 | anisole (l) | 3539 | butyl phenyl ether (l) |
| 5749 | <i>o</i> -propyltoluene (l) | 3365 | <i>o</i> -fluorophenol (v) |
| 4821 | <i>o</i> -cresol (l) | 3343 | catechol (v) |
| 3355 | <i>o</i> -chlorophenol (l) | 4807 | 1,2-dichlorobenzene (v) |
| 3397 | <i>o</i> -cresol (v) | 4809 | 1,2-dibromobenzene (s in CS ₂) |
| 3342 | catechol (s) | 4820 | <i>o</i> -cresol (s in CS ₂) |
| 5525 | <i>o</i> -bromoanisole (s in CHCl ₃) | 4833 | 1,2-dimethoxybenzene (v) |
| 3456 | <i>o</i> -methylbenzyl alcohol (s) | 213 | 1-methyl-2-ethylbenzene (l) |
| 5519 | <i>o</i> -methoxyphenol (s in CHCl ₃) | 3416 | <i>o</i> -xylene (v) |
| 3536 | <i>o</i> -butylanisole (l) | | |
| 217 | 1,3-diethylbenzene (l) | 3417 | <i>m</i> -xylene (v) |
| 2407 | 1,3-dichlorobenzene (l) | 214 | 1-ethyl-3-methylbenzene (l) |
| 2387 | <i>m</i> -chlorotoluene (l) | 4810 | 1,3-dibromobenzene (s in CS ₂) |
| 13758 | <i>m</i> -bromotoluene (l) | 4830 | 1,3-dimethoxybenzene (v) |
| 3968 | <i>m</i> -chlorophenol (l) | 4829 | 1,3-dimethoxybenzene (l) |
| 4822 | <i>m</i> -cresol (s in CS ₂) | 5056 | <i>m</i> -ethylphenol (s in CCl ₄) |
| 3418 | <i>m</i> -xylene (v) | 3396 | <i>m</i> -cresol (v) |
| 219 | 1,4-diethylbenzene (l) | 2383 | 1,4-dichlorobenzene (s) |
| 3352 | <i>p</i> -bromochlorobenzene (s in CS ₂) | 3421 | <i>p</i> -xylene (v) |
| 8205 | <i>p</i> -chloroanisole (s in CHCl ₃) | 3422 | <i>p</i> -xylene (s in CS ₂) |
| 2384 | <i>p</i> -chlorotoluene (l) | 5751 | <i>p</i> -propyltoluene (l) |
| 4041 | <i>p</i> -cresol (l) | 268 | 1-methyl-4-isopropylbenzene (l) |
| 4838 | <i>p</i> -methoxyphenol (s) | 5264 | <i>p</i> -ethylphenol (s in CCl ₄) |
| 3341 | 1,4-dihydroxybenzene (v) | 3535 | <i>p</i> -methoxytoluene (l) |
| 3045 | 1,4-dibromobenzene (s) | | |
| 8780 | 2-chloro- <i>m</i> -cresol (s in CCl ₄) | 6069 | 2,6-diethylphenol (l) |
| 5230 | 2,6-di- <i>sec</i> -butylphenol (l) | 6068 | 2,6-diethylphenol (v) |
| 4803 | 1,2,3-trichlorobenzene (s) | 3534 | 6- <i>sec</i> -butyl- <i>o</i> -cresol (l) |
| 4812 | pyrogallol (s) | 6070 | 2,6-di-isopropylphenol (v) |
| 5733 | 6-chloro- <i>o</i> -cresol (s in CS ₂) | 6067 | 2,6-dimethylphenol (v) |
| 9895 | 2,3-dimethylphenol (s in CS ₂) | 5290 | 1-ethyl-2,3-dimethylbenzene (l) |
| 5728 | 2,6-dichlorophenol (s in CS ₂) | 5293 | 2-ethyl-1,3-dimethylbenzene (l) |
| 967 | 3,5-dihydroxytoluene (s) | 5297 | 1,3-di-isopropyl-5-methylbenzene (l) |
| 9899 | 3-ethyl-5-methylphenol (s in CS ₂) | 5296 | 1-isopropyl-3,5-dimethylbenzene (l) |
| 9898 | 3,5-dimethylphenol (s in CS ₂) | 4805 | 1,3,5-tribromobenzene (s) |
| 5687 | 1,3,5-xenol (s) | 5299 | 1,3,5-tributylbenzene (l) |
| 5688 | 1,3,5-xenol (v) | 5298 | 1,3,5-triisopropylbenzene (l) |
| 5295 | 1,3,5-triethylbenzene (l) | 6518 | 1,3,5-trimethoxybenzene (Nujol) |
| 4804 | 1,3,5-trichlorobenzene (s) | 211 | 1,3,5-trimethylbenzene (l) |
| 5727 | 2,4-dichlorophenol (s in CS ₂) | 4824 | 6-chloro- <i>m</i> -cresol (s) |
| 5228 | 4-ethyl-2- <i>sec</i> -butylphenol (l) | 4832 | 3,4-dimethoxyphenol (s) |
| 5229 | 2,4-di- <i>sec</i> -butylphenol (l) | 4798 | 2-methylquinol (s) |
| 5226 | 5-isopropyl- <i>o</i> -cresol (l) | 5732 | 4-chloro- <i>o</i> -cresol (s in CS ₂) |
| 5227 | 4-ethyl-2-isopropylphenol (l) | 5256 | 4-ethyl-2- <i>sec</i> -butylanisole (l) |
| 9896 | 2,5-dimethylphenol (s in CS ₂) | 5719 | 2-bromomethyl-1,4-dimethoxybenzene (s) |
| 9897 | 3,4-dimethylphenol (s in CS ₂) | 5253 | 2-methyl-4- <i>sec</i> -butylanisole (l) |
| 4802 | 1,2,4-trichlorobenzene (l) | 5254 | 4-methyl-2- <i>sec</i> -butylanisole (l) |
| 212 | 1,2,4-trimethylbenzene (l) | 5255 | 4-ethyl-2-isopropylanisole (l) |
| 8781 | 2,6-dichloro- <i>m</i> -cresol (s in CCl ₄) | T7870-0 | 2,3,6-trimethylphenol |
| 5721 | 2-cyanomethyl-3,6-dimethoxytoluene (s) | C3780-3 ^b | 4-chloro-2,3-dimethylphenol |
| 5722 | 2,2'-bromoethyl-3,6-dimethoxytoluene (s) | 13,698-0 ^b | 2,4,6-trimethylbenzyl chloride |
| 15,360-5 ^b | 1,2,3,4-tetramethylbenzene | 15,347-8 ^b | 2,3,4-trichlorophenol |
| 13,184-9 ^b | 1,2,3,4-tetrachlorobenzene | 15,158-0 ^b | 2,3,6-trichlorophenol |
| T1163-0 ^b | 1,2,3,4-tetrafluorobenzene | 19,586-3 ^b | 2,3,4-trimethoxybenzyl alcohol |
| 15,736-8 ^b | 2,3,4-trichloroanisole | D12,562-8 ^b | 3,6-diisopropylcatechol |
| 13,737-6 ^b | 2,3,6-trichloroanisole | | |
| 13,881-9 ^b | 2,4,6-trimethoxytoluene | T5530-111 ^b | 2,4,6-trichlorophenol |
| 13,878-9 ^b | 1-ethyl-2,4,6-trimethoxybenzene | 15,551-9 ^b | 3,4,5-trichlorophenol |
| 13,877-0 ^b | 2,4,6-trimethoxydiphenylmethane | 19,785-8 ^b | 3,4,5-trimethoxyphenol |
| T7920-0 ^b | 3,4,5-trimethylphenol | D12,560-1 ^b | 3,5-diisopropylcatechol |
| C3830-3 ^b | 4-chloro-3,5-dimethylphenol | D4580-0 ^b | 3,5-di- <i>tert</i> -butylcatechol |
| B6420-2 ^b | 4-bromo-3,5-dimethylphenol | T4940-9 ^b | 2,4,6-Tri- <i>tert</i> -butylphenol |
| T7009-9 ^b | 3,4,5-trimethoxybenzyl alcohol | C3820-6 ^b | 4-chloro-2,6-dimethylphenol |
| 12,599-7 ^b | 2,4-dichloro-6-methylphenol | T1958-5 ^b | 1,2,3,5-tetramethylbenzene |
| T7900-6 ^b | 2,4,6-trimethylphenol | 5731 | 4,6-dichloro- <i>o</i> -cresol (s in CS ₂) |
| T7860-3 ^b | 2,3,5-trimethylphenol | 5691 | mesitol (v) |
| 5783 | 4-chloro-3,5-xenol (Nujol) | 5690a | mesitol (l) |
| 8778 | 2,4,6-trichlorophenol (s in CS ₂) | T1164-9 ^b | 1,2,3,5-tetrafluorobenzene |
| 5232 | 4,6-di- <i>sec</i> -butyl- <i>o</i> -cresol (l) | 15,348-6 ^b | 1,2,3,5-tetrachlorobenzene |
| 5233 | 2,6-di- <i>sec</i> -butyl- <i>p</i> -cresol (l) | | |

Table IX (Continued)

| no. ^a | compd | no. | compd |
|-----------------------|--|-----------------------|--|
| 8777 | 2,4,5-trichlorophenol (s in CS ₂) | 14,674-9 ^b | 4,6-di- <i>tert</i> -butylresorcinol |
| 4801 | 1,2,4,5-tetrachlorobenzene (s) | 11,599-1 ^b | 2,5-bis(methoxymethyl)- <i>p</i> -xylene |
| 270 | 1,2,4,5-tetramethylbenzene (s in CS ₂) | 13,539-9 ^b | 4,6-diisopropyl-1,3-dimethylbenzene |
| 4835 | durene (s) | T1830-9 ^b | 1,2,4,5-tetraisopropylbenzene |
| 13,707-3 ^b | 2,4,5-trichlorotoluene | 19,048-9 ^b | 2,4,5-trimethoxybenzyl alcohol |
| 21,141-9 ^b | 5-bromo-1,2,4-trimethylbenzene | 10,094-3 ^b | 2-bromo-4,5-dimethylphenol |
| 11,615-7 ^b | 2,5-dibromo- <i>p</i> -xylene | T7880-8 ^b | 2,4,5-trimethylphenol |
| T1165-7 ^b | 1,2,4,5-tetrafluorobenzene | C3800-1 ^b | 4-chloro-2,5-dimethylphenol |
| 13,185-7 ^b | 1,2,4,5-tetrachlorobenzene | D6580-1 ^b | 2,4-dichloro-5-methylphenol |
| D7640-4 ^b | 2,5-dichloro- <i>p</i> -xylene | | |

^a DMS index numbers are used. Recording conditions are abbreviated as follows: v = vapor; l = liquid; s = solid; s in x = solution in solvent x.

^b Compounds taken from Aldrich IR Spectra.³⁰ Aldrich Catalog-Handbook number index is used.

Table X. Prediction Results for Discrimination of 10 Aromatic Substitution Patterns with a 2-Hidden Layer Net (96) → [(31) → (15)] → (10)

| test pattern | % correct answers | no. of test spectra | pattern of incorrect answers |
|--------------|-------------------|---------------------|--|
| mono | 71.4 | 7 | [1,2,3,5]; [1,3] |
| [1,2] | 57.0 | 7 | [1,2,3]; [1,2,4,5]; mono |
| [1,3] | 100.0 | 4 | |
| [1,4] | 100.0 | 4 | |
| [1,2,3] | 20.0 | 5 | [1,2,3,4]; [1,2,4,5]; [1,3]; [1,2,3,5] |
| [1,3,5] | 66.7 | 3 | [1,3] |
| [1,2,4] | 63.0 | 8 | [1,2,3]; [1,2]; [1,2,3,4] |
| [1,2,3,4] | 60.0 | 5 | [1,4]; [1,2,3] |
| [1,2,3,5] | 60.0 | 25 | 2× [1,4]; 2× [1,2,4,5]; 4× [1,2,3,4]; 2× [1,3] |
| [1,2,4,5] | 50.0 | 10 | 3× [1,2,3,5]; [1,2,3,4]; [1,3] |

600 cm⁻¹) improves performance to perfect prediction (100% correct answers) if this part of the IR spectrum is digitized with small intervals of 6.25 cm⁻¹. In the case of trisubstituted aromatics, input design does not considerably effect the overall performance of the net. The same is true for tetrasubstitution. With 61.3% correct answers for input type C, predictions, however, are less reliable than in the class of di- and trisubstituted compounds. This is also reflected by the interpretation guides for spectra of aromatic compounds in conventional IR handbooks. Overtone and combination vibrations (region I) and the fingerprint vibrations (region II) of tetrasubstituted aromatics are not very characteristic (Figures 1 and 2).

Effect of Network Architecture on Prediction Performance.

The number of hidden layers necessary for successful pattern recognition depends on the complexity of the problem. Perceptrons (no hidden layers) can only represent linearly separable functions¹⁶ (e.g., OR, AND). Similar input patterns are directly mapped to similar output patterns without any internal representation. One hidden layer is sufficient to approximate any continuous function (e.g., XOR function).¹⁷ An advantage of the perceptron is the simple internal representation of the adapted rules, which can be interpreted and directly visualized as a "learned" pattern in the form of the weight vectors (the contribution of each input unit to a particular output unit, see Figures 5–8). More powerful in solving difficult discrimination problems but less intuitively interpretable are neural nets with hidden layers. Another important parameter in network architecture is the number of neurons in each layer. A minimum number of neurons is necessary to represent the information; an excessive number may allow the net to become overspecific, approximating a look-up table.¹⁴ The optimum number of neurons in our study was found in numerous tests by trial and error.

Tables V–VII show the network performance as a function of the number of hidden layers within each substitution class

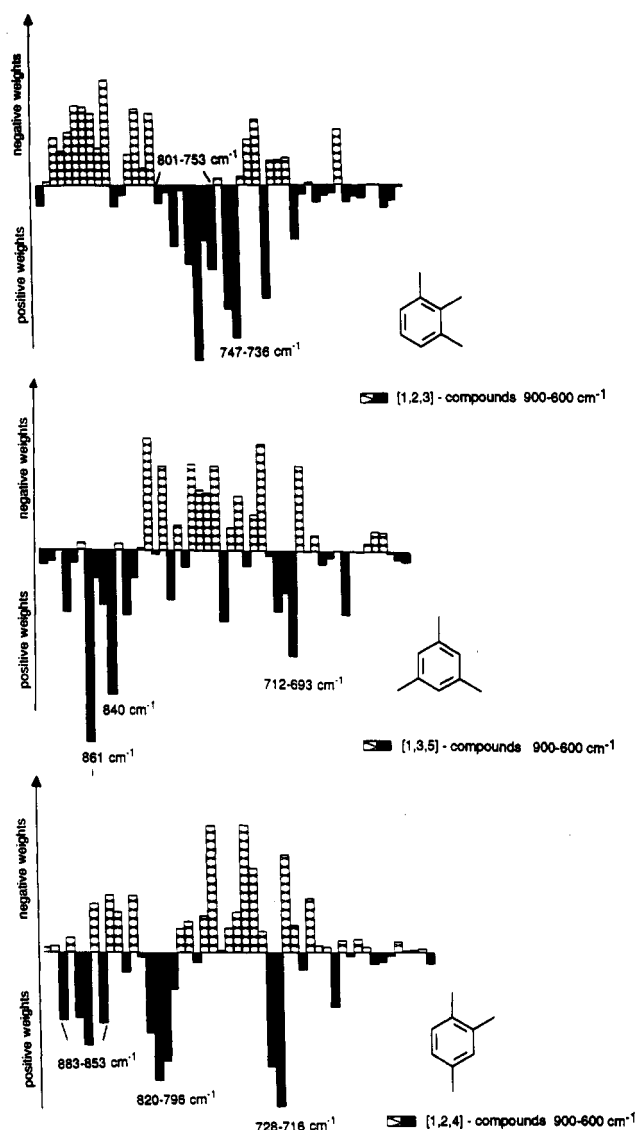


Figure 7. Plot of the weight matrix for region II (fingerprint region) for trisubstituted compounds ([1,2,3], [1,3,5], [1,2,4]). Solid bars represent positive weights, and hatched bars represent negative weights.

(di-, tri-, and tetrasubstituted compounds). The number of correct answers does not considerably improve with an increasing number of hidden layers. The discrimination problem seems to be linearly separable. In Figures 5–8 the weight vectors of the net after the training procedure in the di-, tri-, and tetrasubstituted case are plotted. The weight vectors indicate that the "learned patterns" (solid bars) of IR region I are not very characteristic and of minor importance for interpretation. This is in contrast to the rules traditionally

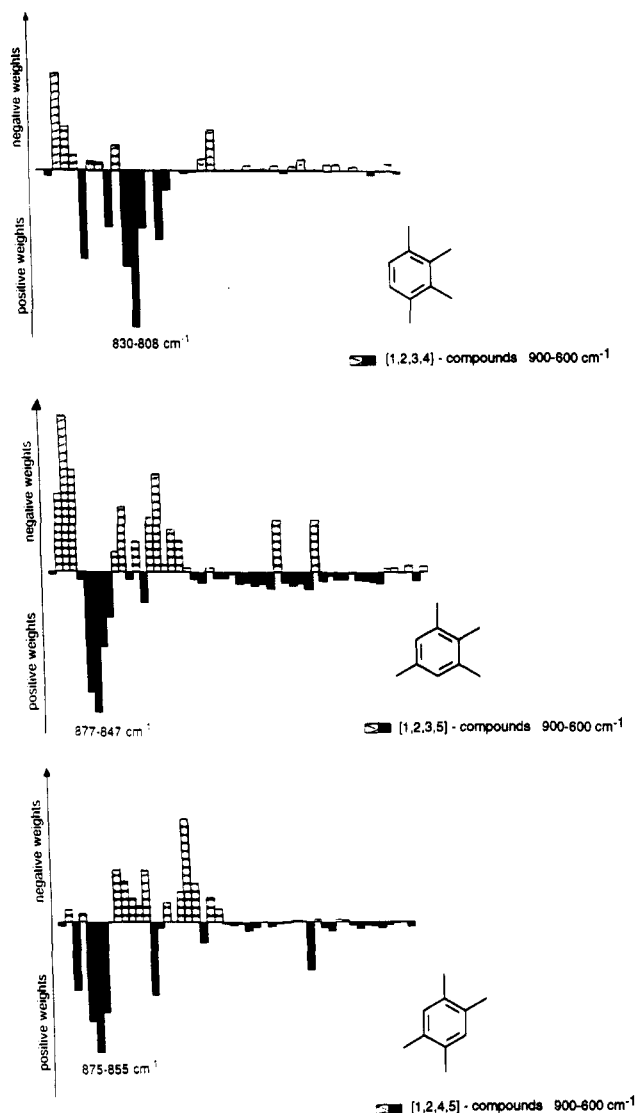


Figure 8. Plot of the weight vector for region II (fingerprint region) for tetrasubstituted compounds ([1,2,3,4], [1,2,3,5], [1,2,4,5]). Solid bars represent positive weights, and hatched bars represent negative weights.

given in IR handbooks¹² (Figure 1) and is probably due to the insufficient quality of our test and training spectra.

Very specific and almost identical with the traditional interpretation rules are the weight vectors in the fingerprint region (region II, Figures 6–8). The net obviously was able to infer the characteristic absorption bands for each substitution pattern by learning from examples. However additional bands were found which so far have not been used in IR spectrum interpretation. Noteworthy is an additional and obviously characteristic band for meta-substituted compounds at 887–857 cm^{-1} , which is not mentioned in IR literature. Another small but characteristic band is found in para-substituted compounds at 650–615 cm^{-1} , which is most probably due to the in-plane ring deformation vibration of the benzene ring.¹⁸ The negative weights are plotted in reverse direction (hatched bars). They can be interpreted as being a negative indication for the corresponding substitution pattern. A band at 760–740 cm^{-1} for instance disfavors the assignment of the spectrum to a meta-substituted compound (Figure 6).

Theoretical Calculations of Vibrational Frequencies. Using the semiempirical MNDO Hamiltonian (MOPAC²⁸ software version 6.0), we calculated the harmonic vibrational frequencies (keyword: FORCE) of mono-, [1,2]-, [1,3]-, [1,4]-,

[1,2,3]-, [1,3,5]-, [1,2,4]-, [1,2,3,4]-, [1,2,3,5]-, and [1,2,4,5]-chlorinated aromatics. A detailed quantum theoretical vibrational analyses of the fingerprint region should provide information about type (normal coordinate analysis) and intensity (transition dipole) of the characteristic fingerprint absorption bands. Thus, the assignment of the characteristic absorptions (weight vectors, see Figures 5–8) to physically meaningful IR vibration modes should be possible. Table VIII shows the computed H out-of-plane wagging vibrations in the fingerprint region (900–600 cm^{-1}) which can be divided into two groups:

- Group 1 Vibrations which involve only the H atoms.
- Group 2 Vibrations which involve the C atoms of the benzene skeleton. (They are present only in mono-, [1,3]-, [1,2,3]-, [1,3,5]-, and [1,2,4]-substituted aromatics.)

The computed frequencies on the average are shifted to 100 cm^{-1} higher wavelengths compared to the experimental absorptions. The discrepancy between the computed and experimental force constants, however, is sufficiently systematic to be corrected by simple scaling procedures.²⁹ The results of the vibrational analysis (Table VIII) compare well with the positive weights of region II (fingerprint region, Figures 6–8) of the weight vector analysis for our nets with no hidden layer.

Studies with Complete Set of Substitution Classes. The overlap of characteristic bands within each substitution class is fairly low, and the pattern recognition problem, as indicated above, was reasonably solved as a linear problem without hidden layers. Considerably more complex is the discrimination between the set of 10 substitution patterns including mono-, di-, tri-, and tetrasubstituted aromatics. Attempts to program networks without hidden layers already failed in the training stage by lack of convergence. The best result was obtained with two hidden layers (96) \rightarrow [(31) \rightarrow (15)] \rightarrow (10) with 62% correct answers after 405 training cycles. Training and test spectra were identical with those in previous studies (Tables III and IX). Table X lists the proportion of correct answers for each substitution pattern. Performance is reasonably good for most classes; however, it fails completely in the case of [1,2,3]-substituted compounds. Spectra of this class were confused with other di- or tetrasubstituted aromatics.

CONCLUSIONS

We have shown that simple neural nets without hidden layers can be trained to recognize the substitution patterns of di-, tri-, and tetrasubstituted aromatic compounds with high (93% for di- and 87% for trisubstitution) to moderate (61.3% tetrasubstitution) accuracy simply by presenting examples of spectra. Performance was comparable or even superior to human experts.²⁵ The characteristic and well-known fingerprint absorptions were reproduced; however, additional bands that are not listed in IR handbooks were autonomously detected as well and used for spectrum interpretation. Discrimination of the set of 10 substitution patterns of mono-, di-, tri-, and tetrasubstituted aromatic compounds required more sophisticated nets with two hidden layers. Obviously, the problem is not linearly separable because of the multiple overlap of characteristic patterns. With 62% correct answers, performance was moderate.

Artificial neural networks are competing with expert system type programs (and of course human experts) in areas where results cannot be obtained from first principles (e.g., quantum mechanics) or by applying simple rules. A disadvantage of

expert systems is the fact that the problem-solving strategy has to be known in advance or extracted from human experts and translated into algorithms. This is sometimes difficult since rules often are applied intuitively and decisions are taken without the awareness of their heuristics (realization). Neural nets, however, learn from examples. In areas where the rules are scarce but the number of examples is large, they seem to be superior to expert systems. Molecular spectroscopy is such an area. The current trend to build up large and easily accessible databases²⁶ will surely benefit applications of neural nets. Neural nets are suitable to collaborate with databases finding spectrum structure correlations as well as enhancing the retrieval of information.²⁷

IR spectrum interpretation using neural nets is probably restricted to functional group recognition^{6,7} and the solution of specific and well-defined structural problems; however, it may be a useful part of computer-assisted structure elucidation systems.

REFERENCES AND NOTES

- (1) (a) Venkatasubramanian, V.; Chan, K. *AIChE J.* **1989**, *35*, 1993. (b) Watanabe, K.; Matsuura, I.; Abe, M.; Kubota, M.; Himmelblau, D. M. *AIChE J.* **1989**, *35*, 1803. (c) Hoskins, J. C.; Himmelblau, D. M. *Comput. Chem. Eng.* **1989**, *12*, 881.
- (2) (a) Bohr, H.; Bohr, J.; Brunak, S.; Cotteril, R. M.; Lautrup, B.; Noershov, L.; Olsen, O. H.; Petersen, S. B. *FEBS Lett.* **1988**, *241*, 223. (b) Qian, N.; Sejnowski, T. J. *J. Mol. Biol.* **1988**, *202*, 865. (c) Holley, L. H.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 152. (d) McGregor, M. J.; Flores, T. P.; Sternberg, M. J. E. *Protein Eng.* **1989**, *2*, 521.
- (3) Lusashin, A. V.; Gragerov, A. I.; Frank-Kamenetskii, M. D. *J. Biomol. Struct. Dyn.* **1989**, *6*, 1123.
- (4) Elrod, D. W.; Maggiora, G. M.; Trenary, R. G. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 477.
- (5) Meyer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. *Science* **1991**, *251*, 542.
- (6) Robb, E. W.; Munk, M. E. *Mikrochim. Acta* **1990**, *1*, 131–155.
- (7) Fessenden, R. J.; Györgyi, L. *Chem. Soc. Perkin Trans. 2* **1991**, 1755.
- (8) Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1–30.
- (9) Carpenter, G. A.; Grossberg, S. The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer* **1988**, *21*, 77–88.
- (10) Kohonen, T. *Self-Organizing and Associative Memory*; Series in Information Sciences; Springer Verlag: Berlin, Heidelberg, New York, and Tokyo, 1984; Vol. 8.
- (11) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing, Vol. 1, Foundations*; MIT: Cambridge, MA, 1988.
- (12) For references see, for example: (a) Weidlein, J.; Müller, U.; Dehnicke, K. *Schwingungsspektroskopie*; Georg Thieme Verlag: Stuttgart, New York, 1988; pp 152–156. (b) Bellamy, G. J. *The Infrared Spectra of Complex Molecules*; John Wiley: New York, 1958. (c) Nakanishi, K. *Infrared Absorption Spectroscopy*; Holden-Day Inc.: San Francisco, 1962. (d) Volkmann, H. *Handbuch der Infrarot-Spektroskopie*; Verlag Chemie: Weinheim, Germany, 1972. (e) Conley, R. T. *Infrared Spectroscopy*; Allyn and Bacon Inc.: Boston, 1966; pp 105.
- (13) Scientific Application International Cooperation: ANSIM.
- (14) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Lecture Notes; Addison-Wesley: Reading, MA, 1991; Vol. 1.
- (15) Source of IR-spectra: DMS spectral cards. *Dokumentation of Molecular Spectroscopy*; Edited in association with Institut für Spektrochemie und Angewandte Spektroskopie Dortmund and DMS Scientific Advisory Board, London: Verlag Chemie: Weinheim/Bergstrasse and Butterworth & Co. Publishers Ltd.: London, 1956.
- (16) Minsky, M.; Papert, S.; *Perceptrons*; MIT Press: Cambridge, MA, 1969.
- (17) Applied to our problem, a simple Perceptron would be able to infer that a spectrum corresponds to a meta-substituted compound, if there is an absorption band at 800 and 700 cm⁻¹, respectively, at 800 or 700 cm⁻¹; however, it could not tackle the problem that if there is a band at 800, an absorption at 880 cm⁻¹ should be absent (XOR), because this would be an indication for a 1,2,4-trisubstituted compound.
- (18) Jakobsen, R. J.; Bentley, F. F. *Appl. Spectrosc.* **1964**, *18*, 88.
- (19) McCulloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133.
- (20) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2554–2558.
- (21) Hopfield, J. J. Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *81*, 3088–3092.
- (22) Hopfield, J. J.; Tank, D. W. Computing with Neural Circuits: A Model. *Science* **1986**, *233*, 625–633.
- (23) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536.
- (24) Some other network models normalize the output to a value between 0 and 1.
- (25) In order to compare our results with human expertise, we presented the same training and test spectra to six students, who in addition had access to a number of IR handbooks and large spectrum collections. Human performance compares well with our net in discriminating ortho-, meta- and para-disubstituted species (87% correct answers versus 93% of the neural net) and is somewhat lower (74% versus 87%) in the case of trisubstituted compounds. Superiority of the program is significant while discriminating the set of 10 substitution patterns of mono-, di-, tri- and tetrasubstituted compounds (62% versus 15%). For a similar comparison between human and artificial neural net performance see: Elrod, D. W.; Maggiora, G. M.; Trenary, R. G. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 466.
- (26) For example SPECINFO. *A Database for IR and NMR of different nuclei*; Chemical Concepts GmbH: Weinheim.
- (27) Meyer, B.; Hansen, T.; Nute, D.; Albersheim, P.; Darvill, A.; York, W.; Sellers, J. *Science* **1991**, *251*, 542.
- (28) MOPAC Software version 6.0: (a) Stewart, J. J. P. *QCPE* 455. (b) Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
- (29) Fogarasi, G.; Pulay, P. *Vibrational Spectra and Structure*; Durig, J. R., Ed.; Elsevier: New York, 1985; Vol. 14, p 125.
- (30) Pouchert, C. J. *The Aldrich Library of Infrared Spectra*, 3rd ed.; Aldrich Chemical Co.: Milwaukee, WI, 1981.