J. CHEM. SOC. FARADAY TRANS., 1994, 90(14), 2015–2019



# Use of a Neural Network to determine the Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals and their Sulfur Analogues

**Driss Cherqaoui, Didier Villemin* and Abdelhalim Mesbah**
*Ecole Nationale Supérieure d'Ingénieurs de Caen (E.N.S.I. de Caen), I.S.M.R.A., U.R.A. 480 CNRS, 6 boulevard du Maréchal Juin, 14050 Caen Cedex, France*

**Jean-Michel Cense**
*Ecole Nationale Supérieure de Chimie de Paris, 11 rue P. et M. Curie, 75005 Paris, France*

**Vladimir Kvasnicka**
*Department of Mathematics, Faculty of Chemical Technology, Slovak Technical University, 81237 Bratislava, Slovakia*



Models of relationships between structure and boiling point (bp) of 185 acyclic ethers, peroxides, acetals and their sulfur analogues have been constructed by means of a multilayer neural network (NN) using the back-propagation algorithm. The ability of a neural network to predict the boiling point of acyclic molecules containing polar atoms is outlined. The usefulness of the so-called embedding frequencies for the characterization of chemical structures in quantitative structure–property studies has been shown. NNs proved to give better results than multiple linear regression and other models in the literature.

NNs have recently[1,2] become the focus of much attention, largely owing to their wide range of applicability and the ease with which they can handle complex and non-linear problems. A leading reference book[3] on the application and the meaning of NN in chemistry has recently been published in which an extensive list of references can be found. NNs have been applied to the identification of proton-NMR spectra,[4] to the interpretation of IR spectra,[5,6] to the prediction of $^{13}C$ chemical shifts,[7] to the classification of mass spectra,[8] to the estimation of aqueous solubilities,[9] to the determination of protein structure,[10,11] to the investigation of quantitative structure–activity relationships (QSAR)[12–14] and to the prediction of chemical reactivity.[15,16]

Boiling point (bp) is one of the properties used to characterize organic compounds. However, it may happen that this property is not available in the literature or difficult to evaluate experimentally. It appears obvious that the usefulness of quantitative structure–property relationships (QSPR) cannot be denied in those cases. Several methods[17,18] for prediction of the bp of organic compounds have been described in the literature. We have recently used NNs to predict the bp of alkanes.[19] These compounds were chosen because they are simple, easy to code and do not have polarized atoms nor intramolecular bonds.

The goals of the current work are: (*a*) To provide an application of the NN theory (developed in our earlier paper[19]) to acyclic ethers, peroxides, acetals and their sulfur analogues.

(*b*) To show the NNs ability to predict the bp of acyclic molecules containing heteroatoms.

(*c*) To call attention to the interest of molecular descriptors such as the embedding frequencies in the presence of heteroatoms.

(*d*) To compare the results obtained by an NN to those given by multiple linear regression (MLR) and to those given in the literature.

## Neural Networks

Artificial NNs are mathematical models of biological neural systems. Three components constitute an NN: the processing elements, the topology of the connections between the nodes (vertices),[20] and the learning rule. In this paper, the specific algorithm used is the back-propagation (BP) system. Its goal is to minimize an error function. A description of the BP algorithm was given previously[19] with a simple example of application and a more extensive description can be found in other works.[21,22]

## Embedding Frequencies

In a BP NN the input layer contains information concerning the data samples under study. In chemistry this information is represented by molecular codes (molecular descriptors). In our study the molecular codes correspond to the embedding frequencies.[23] These integer entities determine to some extent the structure of acyclic compounds composed of carbon, oxygen and sulfur atoms (hydrogen atoms are ignored). Their simple graph-theoretical construction has been described in our recent publication.[24] Let $T$ be a tree with vertices evaluated by symbols C, O or S. $T$ is assigned to any acyclic molecule with skeleton composed of carbon, oxygen, and sulfur atoms. Let $T'$ be a subtree of the tree $T$. $T'$ corresponds to a connected cluster of atoms. The embedding frequency[24,25] of $T'$ in $T$, denoted by $n(T, T')$, is then defined as the number of appearance of the cluster $T'$ in the tree-molecule $T$. In Table 1, 20 clusters used in the construction of input activities, are listed. The input activities correspond to 20 embedding frequencies assigned to these clusters, $d_i = n(T, T_i)$, for $i = 1, 2, \ldots, 20$, where $T$ formally treated as a tree corresponds to a molecule determined by these 20 descriptors. Examples of descriptors for three molecules are listed in Table 2.

**Table 1** List of 20 clusters used for the construction of embedding frequencies

| no. | cluster | no. | cluster | no. | cluster |
|---|---|---|---|---|---|
| 1 | C | 2 | O | 3 | S |
| 4 | C–C | 5 | C–O | 6 | C–S |
| 7 | O–O | 8 | S–S | 9 | C–C–C |
| 10 | C–C–O | 11 | C–O–C | 12 | O–C–O |
| 13 | C–C–S | 14 | C–S–C | 15 | S–C–S |
| 16 | C–C–C–C | 17 | C–(C)3[a] | 18 | C–C–C–C–C |
| 19 | C–C–(C)3[a] | 20 | C–(C)4[a] | | |

[a] 17: isobutyl; 19: isopentyl; 20: neopentyl.

**Table 2** Three examples of 20 descriptors assigned to acyclic molecules

| molecule | $d_1$ $d_{11}$ | $d_2$ $d_{12}$ | $d_3$ $d_{13}$ | $d_4$ $d_{14}$ | $d_5$ $d_{15}$ | $d_6$ $d_{16}$ | $d_7$ $d_{17}$ | $d_8$ $d_{18}$ | $d_9$ $d_{19}$ | $d_{10}$ $d_{20}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| dimethyl peroxide | 2 0 | 2 0 | 0 0 | 0 0 | 2 0 | 0 0 | 1 0 | 0 0 | 0 0 | 0 0 |
| dipropyl sulfide | 6 0 | 0 0 | 1 2 | 4 1 | 0 0 | 2 0 | 0 0 | 0 0 | 2 0 | 0 0 |
| dibutyl disulfide | 8 0 | 0 0 | 2 2 | 6 0 | 0 0 | 2 2 | 0 0 | 1 0 | 4 0 | 0 0 |

## Method

The set of 185 compounds (Table 3) used in the present paper has been studied by Balaban *et al.*[26] This set essentially consists of two basic types of molecules: (1) Acyclic ethers, peroxides and acetals (73 ethers, 17 diethers, 21 acetals and 6 peroxides). (2) Acyclic sulfide, disulfide and thioacetal (45 sulfides, 6 bis-sulfides, 4 thioacetals and 13 disulfides).

**Table 3** Compounds studied with their experimental (exp) bps, predicted (pred) bps and corresponding residuals (res) (all in °C)

| no. | name | $bp_{exp}$ | $bp_{pred}$ | res |
|---|---|---|---|---|
| 1 | dimethyl ether | −23.70 | −4.80 | −18.90 |
| 2 | dimethyl peroxide | 14.00 | 9.81 | 4.19 |
| 3 | dimethyl sulfide | 37.30 | 40.64 | −3.34 |
| 4 | dimethyl disulfide | 109.70 | 112.31 | −2.61 |
| 5 | ethyl methyl ether | 10.80 | 7.50 | 3.30 |
| 6 | ethyl methyl peroxide | 39.00 | 39.24 | −0.24 |
| 7 | dimethoxymethane | 42.00 | 36.41 | 5.59 |
| 8 | ethyl methyl sulfide | 66.60 | 66.95 | −0.35 |
| 9 | ethyl methyl disulfide | 135.00 | 134.93 | 0.07 |
| 10 | bis(methylthio)methane | 148.50 | 150.48 | −1.98 |
| 11 | methyl propyl ether | 40.00 | 35.63 | 4.37 |
| 12 | diethyl ether | 34.60 | 34.98 | −0.38 |
| 13 | isopropyl methyl ether | 32.00 | 31.41 | 0.59 |
| 14 | diethyl peroxide | 63.00 | 58.15 | 4.85 |
| 15 | isopropyl methyl peroxide | 53.50 | 59.66 | −6.16 |
| 16 | ethoxymethoxyethane | 67.00 | 69.19 | −2.19 |
| 17 | 1,1-dimethoxyethane | 64.40 | 64.48 | −0.08 |
| 18 | 1,2-dimethoxyethane | 84.70 | 74.53 | 10.17 |
| 19 | methyl propyl sulfide | 95.50 | 94.82 | 0.68 |
| 20 | diethyl sulfide | 92.00 | 90.89 | 1.11 |
| 21 | isopropyl methyl sulfide | 84.40 | 88.01 | −3.61 |
| 22 | diethyl disulfide | 154.00 | 152.98 | 1.02 |
| 23 | 1,1-bis(methylthio)ethane | 156.00 | 152.97 | 3.03 |
| 24 | ethylthiomethylthiomethane | 166.00 | 167.16 | −1.16 |
| 25 | 1,2-bis(methylthio)ethane | 183.00 | 187.30 | −4.30 |
| 26 | butyl methyl ether | 70.30 | 72.38 | −2.08 |
| 27 | ethyl propyl ether | 63.60 | 62.14 | 1.46 |
| 28 | ethyl isopropyl ether | 52.50 | 54.75 | −2.25 |
| 29 | isobutyl methyl ether | 59.00 | 61.79 | −2.79 |
| 30 | *sec*-butyl methyl ether | 59.50 | 65.29 | −5.79 |
| 31 | *tert*-butyl methyl ether | 55.20 | 53.56 | 1.64 |
| 32 | diethoxymethane | 88.00 | 92.65 | −4.65 |
| 33 | 2,2-dimethoxypropane | 83.00 | 77.85 | 5.15 |
| 34 | 1,3-dimethoxypropane | 104.50 | 105.09 | −0.59 |
| 35 | 1-ethoxy-2-methoxyethane | 102.00 | 104.25 | −2.25 |
| 36 | 1,2-dimethoxypropane | 92.00 | 99.99 | −7.99 |
| 37 | ethyl isopropyl sulfide | 107.40 | 106.47 | 0.93 |
| 38 | butyl methyl sulfide | 123.20 | 124.26 | −1.06 |
| 39 | isobutyl methyl sulfide | 112.50 | 114.89 | −2.39 |
| 40 | ethyl propyl sulfide | 118.50 | 116.96 | 1.54 |
| 41 | *tert*-butyl methyl sulfide | 101.50 | 102.34 | −0.84 |
| 42 | ethyl propyl disulfide | 173.70 | 174.22 | −0.52 |
| 43 | ethyl isopropyl disulfide | 165.50 | 165.57 | −0.07 |
| 44 | bis(ethylthio)methane | 181.00 | 183.98 | −2.98 |
| 45 | methyl pentyl ether | 99.50 | 97.25 | 2.25 |
| 46 | ethyl butyl ether | 92.30 | 93.88 | −1.58 |
| 47 | dipropyl ether | 90.10 | 89.07 | 1.03 |
| 48 | isopropyl propyl ether | 80.20 | 79.21 | 0.99 |
| 49 | ethyl isobutyl ether | 82.00 | 84.03 | −2.03 |
| 50 | isopentyl methyl ether | 91.20 | 86.77 | 4.43 |
| 51 | methyl 2-methylbutyl ether | 91.50 | 87.01 | 4.49 |
| 52 | ethyl *sec*-butyl ether | 81.20 | 83.83 | −2.63 |
| 53 | methyl 1-methylbutyl ether | 93.00 | 85.81 | 7.19 |
| 54 | diisopropyl ether | 69.00 | 69.74 | −0.74 |

**Table 3** (*continued*)

| no. | name | $bp_{exp}$ | $bp_{pred}$ | res |
|---|---|---|---|---|
| 55 | methyl *tert*-pentyl ether | 86.30 | 80.03 | 6.27 |
| 56 | 1,2-dimethylpropyl methyl ether | 82.00 | 85.08 | −3.08 |
| 57 | 1,1-diethoxyethane | 103.00 | 103.38 | −0.38 |
| 58 | 1,1-dimethoxy-2-methylpropane | 103.50 | 103.49 | 0.01 |
| 59 | 2-ethoxy-2-methoxypropane | 96.00 | 96.67 | −0.67 |
| 60 | 1,1-dimethoxybutane | 112.00 | 114.82 | −2.82 |
| 61 | 1-methoxy-1-propoxyethane | 104.00 | 107.02 | −3.02 |
| 62 | 1,4-dimethoxybutane | 132.50 | 131.04 | 1.46 |
| 63 | 1,2-diethoxyethane | 123.50 | 120.49 | 3.01 |
| 64 | 1,3-dimethoxybutane | 120.30 | 123.54 | −3.24 |
| 65 | methyl pentyl sulfide | 145.00 | 146.94 | −1.94 |
| 66 | butyl ethyl sulfide | 144.20 | 143.05 | 1.15 |
| 67 | dipropyl sulfide | 142.80 | 142.02 | 0.78 |
| 68 | isopropyl propyl sulfide | 132.00 | 131.01 | 0.99 |
| 69 | ethyl isobutyl sulfide | 134.20 | 132.84 | 1.36 |
| 70 | isopentyl methyl sulfide | 137.00 | 138.59 | −1.59 |
| 71 | methyl 2-methylbutyl sulfide | 139.00 | 138.21 | 0.79 |
| 72 | *sec*-butyl ethyl sulfide | 133.60 | 131.66 | 1.94 |
| 73 | *tert*-butyl ethyl sulfide | 120.40 | 116.57 | 3.83 |
| 74 | diisopropyl sulfide | 120.00 | 119.84 | 0.16 |
| 75 | 1-ethylpropyl methyl sulfide | 137.00 | 135.40 | 1.60 |
| 76 | dipropyl disulfide | 195.80 | 191.77 | 4.03 |
| 77 | diisopropyl disulfide | 177.20 | 176.05 | 1.15 |
| 78 | *sec*-butyl ethyl disulfide | 181.00 | 185.93 | −4.93 |
| 79 | isopropyl propyl disulfide | 185.90 | 185.14 | 0.76 |
| 80 | *tert*-butyl ethyl disulfide | 175.70 | 172.93 | 2.77 |
| 81 | 1,1-bis(ethylthio)ethane | 186.00 | 185.68 | 0.32 |
| 82 | 1,2-bis(ethylthio)ethane | 211.00 | 210.96 | 0.04 |
| 83 | hexyl methyl ether | 125.00 | 122.66 | 2.34 |
| 84 | ethyl pentyl ether | 118.00 | 115.99 | 2.01 |
| 85 | butyl propyl ether | 117.10 | 117.97 | −0.87 |
| 86 | butyl isopropyl ether | 107.00 | 106.03 | 0.97 |
| 87 | isobutyl propyl ether | 102.50 | 106.13 | −3.63 |
| 88 | ethyl isopentyl ether | 112.00 | 108.48 | 3.52 |
| 89 | *tert*-butyl propyl ether | 97.40 | 92.68 | 4.72 |
| 90 | 2,2-dimethylpropyl ethyl ether | 91.50 | 97.97 | −6.47 |
| 91 | *tert*-butyl isopropyl ether | 87.60 | 87.88 | −0.28 |
| 92 | ethyl 1-methylbutyl ether | 106.50 | 103.06 | 3.44 |
| 93 | ethyl *tert*-pentyl ether | 101.00 | 98.75 | 2.25 |
| 94 | 1,2-dimethylpropyl ethyl ether | 99.30 | 104.49 | −5.19 |
| 95 | ethyl 1-ethylpropyl ether | 90.00 | 105.45 | −15.45 |
| 96 | dipropoxymethane | 137.00 | 134.89 | 2.11 |
| 97 | 2,2-diethoxypropane | 114.00 | 109.58 | 4.42 |
| 98 | 1-ethoxy-1-propoxyethane | 126.00 | 122.16 | 3.84 |
| 99 | 1,1-diethoxypropane | 124.00 | 122.39 | 1.61 |
| 100 | 1,3-diethoxypropane | 140.50 | 139.22 | 1.28 |
| 101 | 1,5-dimethoxypentane | 157.50 | 152.06 | 5.44 |
| 102 | 1-ethoxy-4-methoxybutane | 146.00 | 146.66 | −0.66 |
| 103 | 1,4-dimethoxypentane | 145.00 | 143.88 | 1.12 |
| 104 | 1,3-dimethoxypentane | 141.00 | 144.78 | −3.78 |
| 105 | hexyl methyl sulfide | 171.00 | 169.07 | 1.93 |
| 106 | butyl propyl sulfide | 166.00 | 166.13 | −0.13 |
| 107 | isobutyl propyl sulfide | 155.00 | 155.18 | −0.18 |
| 108 | isobutyl isopropyl sulfide | 145.00 | 147.67 | −2.67 |
| 109 | ethyl 2-methylbutyl sulfide | 159.00 | 153.68 | 5.32 |
| 110 | *tert*-butyl propyl sulfide | 138.00 | 139.41 | −1.41 |
| 111 | *sec*-butyl isopropyl sulfide | 142.00 | 144.81 | −2.81 |
| 112 | ethyl isopentyl sulfide | 159.00 | 154.27 | 4.73 |
| 113 | butyl isopropyl sulfide | 163.50 | 154.47 | 9.03 |
| 114 | 1,3-bis(ethylthio)propane | 229.50 | 225.03 | 4.47 |
| 115 | dibutyl ether | 142.00 | 142.18 | −0.18 |
| 116 | isopentyl propyl ether | 125.00 | 130.53 | −5.53 |
| 117 | butyl isobutyl ether | 132.00 | 129.77 | 2.23 |
| 118 | butyl *sec*-butyl ether | 130.50 | 130.10 | 0.40 |
| 119 | butyl *tert*-butyl ether | 125.00 | 115.23 | 9.77 |
| 120 | *sec*-butyl isobutyl ether | 122.00 | 122.66 | −0.66 |
| 121 | 1,3-dimethylpentyl methyl ether | 121.00 | 133.12 | −12.12 |
| 122 | diisobutyl ether | 122.20 | 119.50 | 2.70 |
| 123 | isobutyl *tert*-butyl ether | 112.00 | 115.69 | −3.69 |
| 124 | di-*tert*-butyl ether | 106.00 | 113.24 | −7.24 |
| 125 | isopropyl *tert*-pentyl ether | 114.50 | 114.79 | −0.29 |
| 126 | heptyl methyl ether | 151.00 | 148.31 | 2.69 |
| 127 | 1-ethylpropyl propyl ether | 128.50 | 126.77 | 1.73 |
| 128 | di-*tert*-butyl peroxide | 109.50 | 101.08 | 8.42 |
| 129 | 1,1-diisopropoxyethane | 126.00 | 130.91 | −4.91 |
| 130 | 1,1-dipropoxyethane | 147.00 | 141.33 | 5.67 |
| 131 | 1,3-dimethoxyethane | 158.00 | 157.66 | 0.34 |
| 132 | 2,4-dimethoxy-2-methylpentane | 147.00 | 146.48 | 0.52 |
| 133 | 1,4-diethoxybutane | 165.00 | 157.69 | 7.31 |
| 134 | dibutylsulfide | 188.90 | 187.68 | 1.22 |
| 135 | diisobutyl sulfide | 170.00 | 169.06 | 0.94 |
| 136 | butyl isobutyl sulfide | 178.00 | 177.68 | 0.32 |
| 137 | di-*tert*-butyl sulfide | 148.50 | 147.73 | 0.77 |
| 138 | di-*sec*-butyl sulfide | 165.00 | 167.32 | −2.32 |
| 139 | butyl *sec*-butyl sulfide | 177.00 | 177.83 | −0.83 |

**Table 3** (*continued*)

| no. | name | $bp_{exp}$ | $bp_{pred}$ | res |
|---|---|---|---|---|
| 140 | *sec*-butyl isobutyl sulfide | 167.00 | 170.87 | −3.87 |
| 141 | heptyl methyl sulfide | 195.00 | 191.54 | 3.46 |
| 142 | dibutyl disulfide | 226.00 | 225.07 | 0.93 |
| 143 | diisobutyl disulfide | 215.00 | 216.22 | −1.22 |
| 144 | di-*tert*-butyl disulfide | 201.00 | 202.29 | −1.29 |
| 145 | 1,1-bis(isopropylthio)ethane | 205.00 | 215.31 | −10.31 |
| 146 | 1-ethyl-1,3-dimethylbutyl methyl ether | 151.50 | 154.31 | −2.81 |
| 147 | ethyl heptyl ether | 165.50 | 161.94 | 3.56 |
| 148 | butyl isopentyl ether | 157.00 | 151.68 | 5.32 |
| 149 | *tert*-butyl isopentyl ether | 139.00 | 142.02 | −3.02 |
| 150 | butyl pentyl ether | 163.00 | 163.76 | −0.76 |
| 151 | 1,5-dimethylhexyl methyl ether | 153.50 | 155.94 | −2.44 |
| 152 | isobutyl isopentyl ether | 139.00 | 148.07 | −9.07 |
| 153 | methyl 1-methylheptyl ether | 162.00 | 160.11 | 1.89 |
| 154 | methyl octyl ether | 173.00 | 175.51 | −2.51 |
| 155 | 2-ethylhexyl methyl ether | 159.50 | 162.75 | −3.25 |
| 156 | methyl 1,1,4-trimethylpentyl ether | 159.50 | 144.30 | 15.20 |
| 157 | 3,5-dimethylhexyl methyl ether | 155.50 | 165.14 | −9.64 |
| 158 | ethyl 1,1,3-trimethylbutyl ether | 141.00 | 142.65 | −1.65 |
| 159 | *tert*-butyl *tert*-pentyl peroxide | 126.00 | 136.20 | −10.20 |
| 160 | 1,1-dimethoxy-2,2-dimethylpentane | 164.00 | 145.47 | 18.53 |
| 161 | 1,1-diethoxypentane | 163.00 | 175.06 | −12.06 |
| 162 | 1,1-dipropoxypropane | 166.50 | 158.68 | 7.82 |
| 163 | 1,1-diisopropoxypropane | 146.00 | 149.80 | −3.80 |
| 164 | 1,3-dipropoxypropane | 165.00 | 185.08 | −20.08 |
| 165 | 1,3-diisopropoxypropane | 159.00 | 152.69 | 6.81 |
| 166 | ethyl heptyl sulfide | 195.00 | 211.74 | −16.74 |
| 167 | methyl octyl sulfide | 218.00 | 215.22 | 2.78 |
| 168 | bis(butylthio)methane | 250.00 | 257.62 | −7.62 |
| 169 | 2,2-bis(propylthio)propane | 235.00 | 225.10 | 9.90 |
| 170 | ethyl octyl ether | 186.50 | 187.80 | −1.30 |
| 171 | ethyl 1,1,3,3-tetramethylbutyl ether | 156.50 | 161.80 | −5.30 |
| 172 | bis(1-ethylpropyl) ether | 162.00 | 160.46 | 1.54 |
| 173 | bis(1-methylbutyl) ether | 162.00 | 160.46 | 1.54 |
| 174 | butyl 1-methylpropyl ether | 170.00 | 173.67 | −3.67 |
| 175 | diisopentyl ether | 173.20 | 168.25 | 4.95 |
| 176 | dipentyl ether | 186.80 | 185.45 | 1.35 |
| 177 | isopropyl heptyl ether | 173.00 | 172.71 | 0.29 |
| 178 | heptyl propyl ether | 187.00 | 185.97 | 1.03 |
| 179 | isopentyl pentyl ether | 174.00 | 176.40 | −2.40 |
| 180 | methyl 1-methyloctyl ether | 188.50 | 186.00 | 2.50 |
| 181 | di-*tert*-pentyl sulfide | 199.00 | 195.40 | 3.60 |
| 182 | dipentyl sulfide | 228.00 | 227.28 | 0.72 |
| 183 | disopentyl sulfide | 215.00 | 210.35 | 4.65 |
| 184 | isobutyl 4-methylpentyl sulfide | 216.00 | 216.52 | −0.52 |
| 185 | methyl nonyl sulfide | 240.00 | 232.63 | 7.37 |

We used a network with 20 units and a bias in the input layer, a variable hidden layer including bias, and one unit in the output layer. Input and output data were normalized between 0.1 and 0.9. The weights were initialized to random values between −0.5 and +0.5 and no momentum was added. The learning rate was initially set to 1 and was gradually decreased until the error function could no longer be minimized.

All computations were performed on an Iris Indigo (Silicon Graphics) workstation using our own programs, written in C language.

## Results and Discussion

In a BP NN the input and output neurons are known since they present, respectively, the embedding frequencies and the bp of the molecules. Unfortunately, there are neither theoretical results available, nor satisfying empirical rules that would enable us to determine the number of hidden layers and of neurons contained in these layers. However, for most of the applications of NNs to chemistry, one hidden layer seems to be sufficient. For the determination of the number of hidden neurons, we have recently[19] discussed the usefulness of the $\rho$ parameter, defined as:

$$\rho = \frac{\text{number of data point in the training set}}{\text{sum of the number of connections in the NN}}$$

**Table 4** Comparison of standard error of learning (SEL) and correlation coefficient ($R$) of NNs, MLR, eqn. (1), eqn. (2) and eqn. (3)

| method | SEL | $R$ |
|---|---|---|
| NN3 | 3.507 | 0.997 |
| NN4 | 3.311 | 0.998 |
| NN5 | 2.942 | 0.998 |
| NN6 | 2.685 | 0.998 |
| NN7 | 2.800 | 0.998 |
| NN8 | 2.948 | 0.998 |
| MLR | 6.350 | 0.992 |
| eqn. (1) | 9.0 | 0.982 |
| eqn. (2) | 10.5 | 0.977 |
| eqn. (3) | 8.2 | 0.986 |

NN3 ... NN8 is for 3 ... 8 neurons in the hidden layer.

According to Zupan and Gasteiger[27] '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' (*i.e.* $\rho \geqslant 1$). In this paper, six architectures of NN (20–$x$–1; $x$ = 3, 4, 5, 6, 7, 8; *i.e.* $\rho \in [1.05, 2.76]$) have been tried, and two studies have been achieved: learning and prediction. The term learning is used when the NN estimates bp values for molecules in the training set. When it estimates bp values for molecules not included in the training set, this is prediction.

### Learning

*NNs*

In order to determine the best architecture, six different ones have been tried (20–$x$–1; $x$ = 3, 4, 5, 6, 7, 8). The criteria used for the comparison of the six architectures are the correlation coefficient ($R$) and the standard error of learning (SEL) defined by:

$$SEL^2 = \frac{\sum (bp_{exp} - bp_{calc})^2}{N}$$

$$R^2 = 1 - \frac{\sum (bp_{exp} - bp_{calc})^2}{\sum (bp_{exp} - bp_{mean})^2}$$

where $bp_{mean}$ stands for the arithmetic mean of all $N$ observed values of the bp.

The results obtained are given in Table 4. Fig. 1 clearly indicates that the SEL goes down to a minimum corresponding to six neurons in the hidden layer. It can be seen that the SEL increases slightly (*i.e.* the learning performance decreases) for seven and eight neurons. That is due to the fact that the number of weights is nearly equal to the number of molecules in the training set. Thus, the information brought

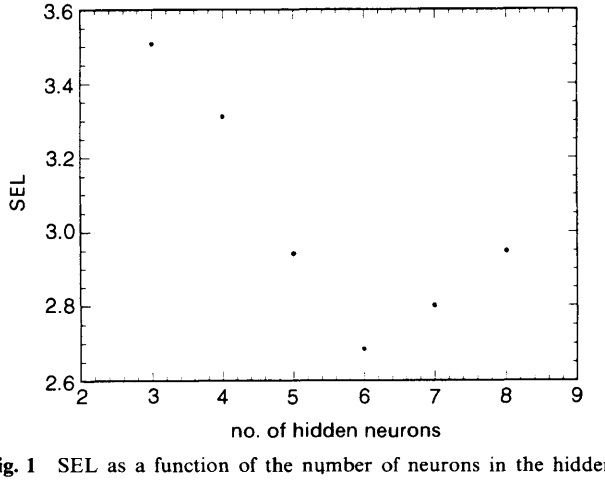


**Fig. 1** SEL as a function of the number of neurons in the hidden layer

to the training set is not sufficient to train correctly the NN with the architecture 20–$x$–1 ($x$ = 7 and 8).

*MLR*

The most widely used mathematical method in QSAR or QSPR is MLR. The objective of such an analysis is to find an equation that relates a dependent variable (such as the bp property) to one or more independent variables (such as molecular descriptors). The solution to the problem consists in determining the coefficients $a_i$ and the constant term $a_0$ of the following equation:

$$\text{bp} = a_0 + \sum a_i d_i$$

It is helpful to note some inherent difficulties of MLR in particular, arising from the interdependence of molecular descriptors. In this study MLR was used to correlate bp with only 15 independent molecular descriptors ($d_6$, $d_7$, $d_8$, $d_{11}$ and $d_{15}$ are removed). The correlation coefficient and the standard error of learning are 0.992 and 6.350, respectively.

*Other Models in the Literature*

Bps of the 185 compounds studied were correlated by Balaban *et al.*[26] with chemical structures using two or three topological descriptors. Three equations were found:

$$\text{bp} = -59.10 + 44.30\,{}^1\chi + 42.88N_s; \quad R = 0.982; \quad S = 9.0 \tag{1}$$

$$\text{bp} = -11.23 - 7.21S_{\text{het}} + 35.04\,{}^0\chi^v - 18.30T_{\text{Me}}; \quad R = 0.977; \quad S = 10.5 \tag{2}$$

$$\text{bp} = -41.75 + 43.79\,{}^1\chi + 45.03N_s - 2.90J_{\text{het}}; \quad R = 0.986; \quad S = 8.2 \tag{3}$$

All the results given by NNs, MLR, eqn. (1), eqn. (2) and eqn. (3) are shown in Table 4. We see that in all cases the NN approach gives the best results. However, the learning abilities of the models are not completely comparable since the descriptors used are not the same. In this study NNs show an interesting ability to extract information about cyclic compounds directly from the embedding frequencies.

*Prediction*

The predictive ability of an NN is its ability to give a satisfying output to a molecule not included in the examples the NN learned. To determine that predictive ability, cross-validation has been used. In this procedure one compound is removed from the data set, the network is trained with the remaining compounds and used to predict the discarded compound. The process is repeated in turn for each compound in the data set. After cross-validation, the predictive ability of different networks was assessed by the standard error of prediction (SEP) and the cross-validated $R^2$ ($R^2_{cv}$).

$$\text{SEP}^2 = \frac{\sum (\text{bp}_{\text{exp}} - \text{bp}_{\text{pred}})^2}{N}$$

$$R^2_{cv} = 1 - \frac{\sum (\text{bp}_{\text{exp}} - \text{bp}_{\text{pred}})^2}{\sum (\text{bp}_{\text{exp}} - \text{bp}_{\text{mean}})^2}$$

**Table 5** Comparison of predictive ability for NNs and MLR

| method | SEP | $R^2_{cv}$ |
|---|---|---|
| NN3 | 5.223 | 0.988 |
| NN4 | 5.152 | 0.988 |
| NN5 | 5.102 | 0.989 |
| NN6 | 5.946 | 0.985 |
| NN7 | 6.215 | 0.983 |
| MLR | 6.710 | 0.981 |

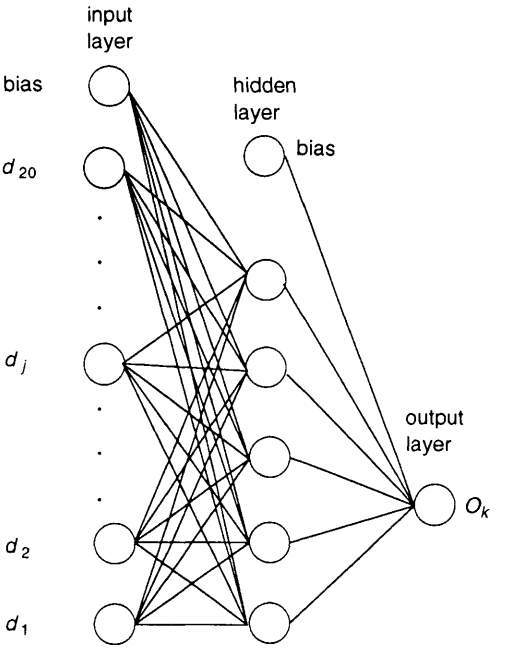


**Fig. 2** Architecture of a BP network with three layers. The configuration shown is 20–5–1.

Table 5 shows the results obtained with five different architectures and with MLR. This table shows that the NN performance is a function of the number of hidden neurons. NNs give a superior performance to that given by MLR. In MLR the relationship between bp and molecular descriptors is expressed by a linear combination of the contributing terms. On the contrary the NN owes its predictive ability to its non-linear power. This does not mean that the NN is a polynomial model but it is able to learn by example how to make predictions for cases not belonging to the training set. It can be seen that the best architecture is 20–5–1 ($\rho$ = 1.67; Fig. 2). It is interesting to note the variation of the SEP according to the number of iterations. Fig. 3 shows this variation for the NN with an architecture 20–5–1. The learning performance of the NN increases with the number of iterations, but its predictive ability slowly decreases after 4000 iterations. This is known as the overtraining effect, due to a too long learning time. Indeed, the weights obtained after the overtraining contain more information specific to the training set. Therefore, prediction on the test set will not really be satisfying. Thus, when a very low error in the training set is

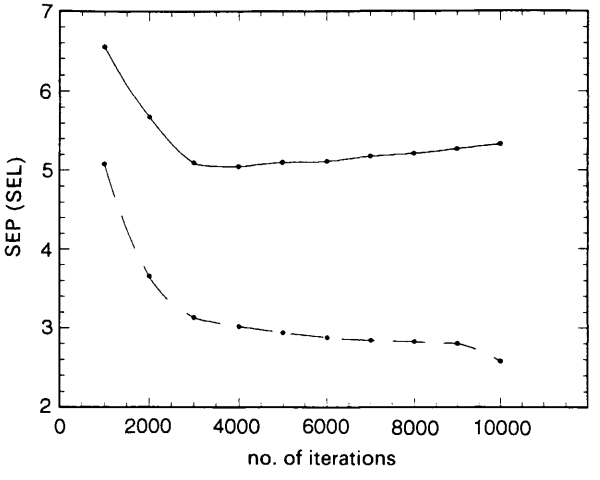


**Fig. 3** Predictive ability of NN (top curve). Learning ability of NN (bottom curve).

sought, the predictive ability of an NN is less successful. The ability to predict being an essential quality of an NN, the overtraining effect must be avoided. The full results of cross-validation for 4000 iterations and with the NN architecture 20–5–1 are gathered in Table 3. Those results are satisfying and show that the embedding frequencies are very useful descriptors for the compounds studied. Nevertheless, six outliers can be seen (compounds 1, 95, 156, 160, 164 and 166 with residuals between 15 and 20 °C). For dimethyl ether, a large deviation is expected because it is the only one to have a negative experimental bp. It should be noted that the NN predicted a negative value for this compound. Since the bp is one of the physical properties that are difficult to measure,[28] the experimental bps of the other outliers may be in error.

## Conclusion

This paper has discussed the use of BP NN to predict the boiling point of acyclic ethers, peroxides, acetals and their sulfur analogues. The performances of NN were compared with those given by MLR and those of other models in the literature, and proved to be better. It is interesting to note that the performances of the NNs decrease when overtraining occurs. The embedding frequencies provide enough information to an NN for prediction of the bp of the compounds studied. The approach using the embedding frequencies is adapted to the modelling of compounds containing heteroatoms, which is not the case for descriptors based on topological indices.[29]

## References


1 J. Zupan and J. Gasteiger, *Anal. Chim. Acta*, 1991, **248**, 1.
2 M. Tusar, J. Zupan and J. Gasteiger, *J. Chim. Phys.*, 1992, **89**, 1517.
3 J. Zupan and J. Gasteiger, in *Neural Networks for Chemists*, VCH, New York, 1993.
4 J. U. Thomsen and B. Meyer, *J. Magn. Reson.*, 1989, **84**, 212.
5 E. W. Robb and M. E. Munk, *Mikrochim Acta (Wien)*, 1990, **I**, 131.
6 M. E. Munk, M. S. Madison and E. W. Robb, *Mikrochim Acta (Wien)*, 1991, **II**, 505.
7 V. Kvasnicka, *J. Math. Chem.*, 1991, **6**, 63.
8 B. Curry and D. E. Rumelhart, *Tetrahedron Comput. Methodol.*, 1990, **3**, 213.
9 N. Bodor, A. Harget and M. J. Huang, *J. Am. Chem. Soc.*, 1991, **113**, 9480.
10 L. H. Holley and M. Karplus, *Proc. Natl. Acad. Sci. USA*, 1989, **86**, 152.
11 N. Qian and T. J. Sejnowski, *J. Mol. Biol.*, 1988, **202**, 865.
12 D. Villemin, D. Cherqaoui and J-M. Cense, *J. Chim. Phys.*, 1993, **90**, 1505.
13 T. Aoyama and H. Ichikawa, *Chem. Pharm. Bull.*, 1991, **39**, 358.
14 T. Aoyama and H. Ichikawa, *Chem. Pharm. Bull.*, 1991, **39**, 372.
15 V. Simon, J. Gasteiger and J. Zupan, *J. Am. Chem. Soc.*, 1993, **115**, 9148.
16 D. W. Elrod, G. M. Maggiora and R. G. Trenary, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 477.
17 D. E. Pearson, *J. Chem. Educ.*, 1957, **28**, 60.
18 R. D. Cramer, *J. Am. Chem. Soc.*, 1980, **102**, 1837.
19 D. Cherqaoui and D. Villemin, *J. Chem. Soc., Faraday Trans.*, 1994, **90**, 97.
20 N. Trinajstic, in *Chemical Graph Theory*, CRC Press, Boca Raton, FL, 1992.
21 J. L. McClelland, D. E. Rumelhart and the PDP Research Group, in *Parallel Distributed Processing*, ed. J. L. McClelland and D. E. Rumelhart, MIT Press, Cambridge, MA, 1988, vol. I, p. 319.
22 J. A. Freeman and D. M. Skapura, in *Neural Networks Algorithms, Applications, and Programming Techniques*, Addison-Wesley, Reading, 1991, p. 89.
23 R. D. Poshusta and M. C. McHugues, *J. Math. Chem.*, 1989, **3**, 193.
24 D. Cherqaoui, D. Villemin and V. Kvasnicka, *Chemom. Intell. Lab. Syst.*, in the press.
25 V. Kvasnicka, D. Cherqaoui and D. Villemin, *J. Comput. Chem.*, in the press.
26 A. T. Balaban, L. B. Kier and N. Joshi, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 237.
27 Ref. 3, p. 263.
28 M. Randic, *Croat. Chem. Acta*, 1993, **66**, 289.
29 M. Randic and N. Trinajstic, *J. Mol. Struct. (Theochem)*, 1993, **284**, 209.


*Paper* 3/07329G; *Received* 13*th December*, 1993