Correlation between Structure and Normal Boiling Points of Haloalkanes C_1–C_4 Using Neural Networks

Alexandru T. Balaban,*,† Subhash C. Basak,‡ Timothy Colburn,§ and Gregory D. Grunwald‡

Department of Organic Chemistry, Polytechnic University, Splaiul Independentei 313, 77206, Bucharest, Roumania, Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth, Minnesota, 55811, and Department of Computer Science, University of Minnesota, 320 Heller Hall, Duluth, Minnesota 55812

Received March 7, 1994®

By using neural networks, correlations were established between chemical structure and boiling points of chlorofluorocarbons with 1, 1–2, or 1–4 carbon atoms (15, 62, and 276 compounds, respectively) as well as of halomethanes with up to four different halogens (48 compounds). The molecular descriptors included the number of carbon atoms and of each type of halogen atom as well as topological indices. Results were validated by the jackknifing procedure. The correlation coefficients were r = 0.985–0.995. Predictions were made for the boiling points of several haloethanes.

INTRODUCTION

The use of neural networks (NNs) in QSAR-QSPR studies[1-13] offers definite advantages over traditional methods when the activity (in QSAR) or property (in QSPR) does not vary linearly with parameters and/or when these parameters have to be identified. Caution must be exercised in order to avoid overtraining and resulting in memorization of input/output patterns; this can be done by submitting data in varying order during the training period of the NN. The validation should be checked by leaving out some of the data during the training period and then, after the training period, by including these data. Finally, by means of the weights associated with the connections of the hidden layer(s), one can identify relevant parameters and return to modeling the QSAR/QSPR data by means of linear or nonlinear procedures such as principal components analysis or regression (PCA, PCR), multiple linear regression (MLR), or nonlinear partial least squares (PLS).[14,15]

In a previous paper,[16] boiling points at normal pressure (BPs) of 530 haloalkanes with 1–4 carbon atoms were correlated with their structure by using molecular descriptors including topological indices and linear equations. It was pointed out, however, that the variation *versus* the number of halogen or carbon atoms was not linear. This was most evident when the halogen was fluorine because with the heavier halogens (Br, I) the marked increase in BPs partially obscures the nonlinearity. Moreover, with bromo and iodo derivatives with known BPs, the number of halogen atoms in the molecule is smaller than for fluoro or chloro derivatives; because of their marked increase of BPs, most derivatives with higher numbers of Br or I atoms decompose on heating at normal pressure, and their BPs have been determined only at lower pressures. Nonlinear equations with maxima situated asymmetrically with respect to the ascending and descending branches of the curve (BP *versus* numbers of halogen atoms) introduce several adjustable parameters which, together with other molecular descriptors, raise considerably the number of parameters to be determined as the essential ones. The previous study[16] led to reasonably good correlation ($r^2 \geq 0.97$) and to modest standard deviations (5.3° for the subset of 44 halomethanes where no isomerism exists and 11° for the entire set of 530 haloalkanes C_1–C_4).

The present paper reports results for correlating chemical constitution with normal BPs for some of these compounds by using NNs.

By applying NNs to the set of compounds having oxygen or sulfur atoms and no hydrogen bonds (for which in our previous study using linear regressions and similar molecular descriptors as for haloalkanes,[17] better r and s values had been obtained), Lohninger[18] was able to further improve considerably the correlation.

NEURAL NETWORK TOOLS AND METHODS

A neural network simulator was used to predict boiling points of compounds based on numbers of carbon and/or halogen atoms and several topological indices. The project proceeded in two stages. In the first stage, artificial neural networks were designed and trained for increasingly larger sets of compounds and with increasing number of inputs in order to validate the neural network model as a predictor of boiling points for compounds whose boiling points are known. In the second stage, the validated model was used to predict the boiling points of compounds whose boiling points are not known.

In the neural network selected for each study, the architecture had three layers, as will be shown in Tables 1 and 2. The number of nodes in the hidden layer was chosen according to a rule of thumb which may be formulated as follows: the more hidden nodes the better, provided that the total number of nodes in the architecture does not exceed the number of input cases. This constraint is intended to keep the set from "memorizing" the data (i.e., from overtraining) which it would have a tendency to do if there are too many connections in which to store weights. Thus the choice of the number of hidden nodes comes down to a trade-off between overlearning and not learning sufficiently.

RESULTS AND DISCUSSION

To validate boiling point data for a set of N compounds whose boiling points are known, we chose a set of M descriptors as inputs to train a neural net. It was important that M be not too large in relation to N to avoid an oversize number of

† Polytechnic University.
‡ Natural Resources Research Institute, University of Minnesota.
§ Department of Computer Science, University of Minnesota.
® Abstract published in *Advance ACS Abstracts,* August 1, 1994.

0095-2338/94/1634-1118$04.50/0

Table 1. Validation Models for BPs Using Neural Networks with M Descriptors and N Haloalkanes with C Carbon Atoms

model	N	M	C	neural net inputs	neural net architecture	std error (°C)	training cycles per compound	correlation coeff of computed and actual BP
1	15	2	1	Cl, F	2-4-1	11.8	40,000	0.985
2	48	4	1	Cl, F, Br, I	4-8-1	7.9	30,000	0.995
3	62	4	1, 2	C, Cl, F, W	4-12-1	8.3	20,000	0.992
4	276	5	1–4	C, Cl, F, W, J	5-10-1	8.5	2,000	0.992

Table 2. Neural Network Architectures and Parameters for Actual BP Models 1 and 2 Which Will Be Used Henceforth

actual model	no. of C atoms	parameters used for training	no. of compds	network architecture	squared error after training
model 1	1, 2	C, F, Cl, Br, I; W, J	171	7-14-1	0.000 065
model 2	2	F, Cl, Br, I; W, J	123	6-12-1	0.000 105

interconnections in the net, leading to overtraining and the resulting memorization of input/output patterns. Validation of the chosen descriptors as modelers of BPs then proceeded as follows: all but one of the N compounds were used to train the net, using backpropagation, learning $N-1$ boiling points to within an accepted margin given by the mean squared error. The trained net was then used, in feed-forward mode only, to compute the boiling point of the one compound held out from the training. This process was repeated for all N compounds, resulting in a vector of N computed boiling points which was then analyzed for correlation with the vector of actual points. Because this procedure, sometimes called the "jackknife" method, is time-consuming for large N values, we began cautiously with $N = 15$ and $M = 2$. As we became more confident in the approach, we repeated the process with increasingly larger values of both N and M. By the end of the validation N and M were 276 and 5, respectively, and the jackknife process for the 60 connections took over 40 h to complete on a Sun-4 workstation. For actual models 1 and 2 to be used below, such a jackknifing would be prohibitively long.

Following is the set of descriptors from which we chose our neural net inputs: C, number of carbon atoms; F, number of fluorine atoms; Cl number of chlorine atoms; Br, number of bromine atoms; I, number of iodine atoms; W, Wiener number (see below);[19] and J, J index (see below).[20]

The Wiener index is the half-sum of all entries in the graph-theoretical distance matrix; the sums over rows or columns i in this matrix are called distasums, S_i; the Wiener and J indexes can also be formulated as follows:

$$W = 1/2\sum_i S_i$$

$$J = \frac{q}{q-p+2} \sum_{\text{edges } ij} (S_i S_j)^{-1/2}$$

where p and q are the numbers of graph vertices and edges, respectively.

The boiling point model validation steps are summarized in Table 1.

It can be seen from Table 1 that in the last three models the correlation coefficient r is higher than 0.99. In the first model, the predicted values led to deviations that were larger than 13° for only four compounds: CH_4, CH_3Cl, CH_3F, and CF_4. In the second model, only two outliers above the previously mentioned limit appeared, namely CH_3F and CF_3I. In the third model, five outliers appeared, namely CF_4, $(FCH_2)_2$, C_2Cl_6, CCl_4, and EtCl. In the last model, the most serious outliers (by 30–32°) were $H_3C–CH_2–CH_2–CCl_3$ and $H_3C–CFCl–CH_2–CF_2Cl$.

Table 3. Predicted BPs for Three Dihaloethanes Using Actual Models

	descriptors							predicted boiling point (°C)		
compd	C	F	Cl	Br	I	W	J	model 1	model 2	ref 1–16
Me—CHBrF	2	1	0	1	0	9	2.324	40.3	37.6	42
Me—CHFI	2	1	0	0	1	9	2.324	68.9	72.1	76
$ICH_2—CH_2I$	2	0	0	0	2	10	1.975	198.6	195.7	196

In the previous study,[16] three predictions were made for boiling points: $H_3C–CHBrF$ (42 °C), $H_3C–CHFI$ (76 °C), and $ICH_2–CH_2I$ (196 °C). In order to compare these predictions with those based on neural networks, two actual models were run as shown in Table 2.

For both models there were 2000 cycles involved in the training. The first 1000 cycles were run with the learning rate at 0.7 and the momentum at 0.9. For the last 1000 cycles these were dropped to 0.2 and 0.4, respectively. For each cycle (or epoch), inputs were chosen for training in a random order.

The boiling points predicted for the three test compounds are given in Table 3.

One can see that the predicted normal boiling points are in good agreement for the two models and for the data originating in the previous study, but for the first two compounds the present correlations predict values which are lower by 3–5 °C.

In a separate study, the number of topological indices was increased to ten by adding eight other ones, namely, K_1, K_4, SIC_0, SIC_3, $S5$, $SPC5$, $V2$, and $VPC4$.

Indices K_1 and K_4 are the numbers of paths of lengths 1 and 4, respectively. Indices SIC_0 and SIC_3 are structural information content indices developed by Basak et al.[21,22] Indices $S5$, $SPC5$, $V2$, and $VPC4$ have been introduced by Kier and Hall and represent the following: the fifth order simple path connectivity index, the fifth order simple path-cluster connectivity index, the second order valence-corrected path connectivity index, and the fourth order valence-corrected path-cluster connectivity index, respectively.[23]

The architecture was 13–18–1. The percentage contributions for the following ten topological indices W, J, K_1, K_4, SIC_0, SIC_3, $S5$, $SPC5$, $V2$, and $VPC4$ are 6.4, 6.0, 10.0, 4.9, 7.0, 8.6, 5.9, 4.5, 8.1, and 5.1, respectively. However, on running two extra models by replacing in the preceding two models the indices W and J by K_1 and SIC_3 (which had the largest input nodes' shares of output layer value), the training errors were higher than in the preceding models (Table 2), and the predicted boiling points did not agree among themselves or with the earlier predictions.[16]

We consider, therefore, that W and J (which are practically not intercorrelated) give the best result in the present context.

In order to make it possible for interested persons to use the NN presented in the present paper, Table 4 presents for model 1 the weights of the connections between the input and hidden

Table 4. Neural Net for Actual Model 1[a]

Hn	C	Cl	F	Br	I	W	J
H1	−0.935 076	−0.972 509	0.305 402	1.088 34	0.143 585	0.175 889	−0.712 527
H2	−0.731 989	−1.463 11	0.252 674	−0.224 324	−1.588 96	−1.311 59	−1.128 57
H3	−2.281 13	2.012 64	−2.290 93	−2.507 31	−1.879 59	0.955 615	1.132 24
H4	−1.145 79	0.365 782	−0.508 501	−0.122 495	−0.005 08383	0.502 795	−0.335 706
H5	0.591 581	−2.637 22	2.200 6	3.147 57	3.080 73	1.599 49	−1.844 85
H6	0.591 581	−0.217 318	0.608 833	−0.374 436	0.198 162	−0.304 568	−0.624 789
H7	−0.795 006	0.498 989	−0.218 393	−0.672 204	−0.192 084	−0.916 662	−1.062 53
H8	−0.499 39	−1.384 22	1.324 99	1.937 73	1.223 01	0.910 384	−1.245 97
H9	0.921 179	−1.007 77	1.721 55	1.456 1	1.014 48	−0.157 602	−2.324 39
H10	−0.038 5723	−2.400 92	−1.984 27	−0.759 809	0.046 5334	−4.166 33	1.191 43
H11	−0.070 4953	0.369 587	−2.231 81	−2.804 06	−3.041 24	−2.036 32	3.330 7
H12	0.084 5828	0.422 513	−0.634 86	0.876 337	0.158 061	−0.707 963	−1.697 99
H13	−2.540 66	1.651 21	−1.552 29	−1.434 61	−1.060 98	−0.244 693	−0.346 334
H14	0.647 746	−2.100 58	−3.317 67	−2.654 78	−1.524 32	−3.674 02	−5.023 83

[a] Connections between the input and hidden layers are described by a weight matrix; the *n*th row (H*n*) of the matrix specifies the weights going into the *n*th hidden node from each of the seven input nodes corresponding to the descriptors (parameters) used for training.

Table 5. Data for the 14 Hidden Nodes of Actual Model 1[a]

node	bias elements	connection wts
H1	−0.556 171	1.017 34
H2	−0.798 285	−1.522 95
H3	−2.953 99	−3.878 96
H4	−1.463 59	−0.171 569
H5	−3.591 92	3.603 16
H6	−1.263 63	0.659 915
H7	−0.877 281	0.438 451
H8	−1.735 06	1.747 23
H9	−0.049 9956	1.137 23
H10	−0.309 85	−2.386 03
H11	−2.606 72	−3.150 08
H12	−1.052 05	1.134 61
H13	−1.669 31	−1.836 87
H14	−0.782 972	−4.343 03

[a] Bias elements for the hidden nodes and connection weights between the hidden and output layer.

Table 6. Neural Net for Actual Model 2[a]

Hn	Cl	F	Br	I	W	J
H1	−0.764 917	−0.521 487	0.261 745	0.978 076	−0.236 504	−0.150 165
H2	−0.899 283	0.132 91	0.966 481	0.504 211	0.185 71	−0.903 144
H3	−2.160 08	−3.655 77	−2.680 25	−1.725 26	−5.637 45	−0.434 807
H4	−1.346 88	−0.433 736	0.123 999	−0.537 107	−1.375 4	−0.639 233
H5	−0.714 541	0.235 642	2.497 73	0.599 368	0.418 764	−1.752 98
H6	−3.055 64	2.63354	3.186 97	2.871 08	1.063 74	−1.193 09
H7	−1.606 48	−1.360 67	−0.434 695	−0.672 228	−2.774 27	0.276 215
H8	−0.695 511	1.620 51	1.408 86	−0.672 228	2.020 82	−3.794 58
H9	−1.008 53	−0.334 537	0.323 877	−0.693 551	−0.971 639	−1.142 75
H10	0.979 414	−2.000 32	−3.228 14	−1.840 21	−1.140 97	1.986 63
H11	−2.010 4	0.317 293	1.355 39	−0.640 517	0.365 577	−0.670 698
H12	2.029 91	−2.558 09	−2.749	−2.350 34	0.410 922	1.225 39

[a] Connections between the input and hidden layer are described by a weight matrix; the *n*th row (H*n*) of the matrix specifies the weights going into the *n*th hidden node from each of the six input nodes corresponding to the descriptors (parameters) used for training.

layers; Table 5 contains the bias elements for the hidden nodes and the connection weights between the hidden and output layers; the bias element for the output node is −0.264 346. Corresponding data for model 2 are given in Tables 6 and 7; the bias element for the output node is −1.564 72.

It must remembered that these weights are those obtained by training nets on input and target values which have been scaled to the interval [0,1] and then unscaled after prediction. Our scaling algorithm depends upon the min. and max values of the input and target values in the training set. For information about these values and the algorithm, please contact tcolburn@d.umn.edu.

Table 7. Data for the 12 Hidden Nodes of Actual Model 2[a]

node	bias elements	connection wts
H1	−1.570 69	0.569 729
H2	−1.720 13	0.846 644
H3	−0.938 863	−5.128 29
H4	−1.052 4	−0.748 751
H5	−2.343 84	2.005 86
H6	−3.241 59	3.241 04
H7	−0.552 587	−2.275 42
H8	1.942 95	2.212 36
H9	−0.950 345	−0.243 132
H10	−2.450 26	−2.946 44
H11	−2.017 81	1.020 06
H12	−4.150 87	−3.725 4

[a] Bias elements for the hidden nodes and connection weights between the hidden and output layer.

CONCLUSIONS

By using neural networks, the preceding correlation between normal boiling points of haloalkanes and their chemical constitution was improved, and reliable predictions for BPs of three dihaloethanes were obtained. Molecular descriptors included the numbers of carbon and/or halogen atoms of different kinds and topological indices.

One should stress that topological indices are valuable because they can be easily calculated for known and unknown structures. They do not discriminate among stereoisomers, but thermal properties differ little between diastereoisomers and not at all among enantiomers.

Special attention was paid to chlorofluorocarbons (CFCs) and hydrogen-containing chlorofluorocarbons (HCFCs) with one or two carbon atoms. The stability and lack of toxicity of CFCs and HCFCs have made them the most useful potential agents to be used as foaming agents in polymer manufacture, in spray cans as propellants, and in heat pumps employed for air conditioners and for refrigerators.

However, the stability of CFCs allows them to reach the stratosphere, where their photochemical generation of chlorine atoms causes destruction of the ozone layer; the Montreal Protocol phases out the manufacture and use of CFCs; by contrast, HCFCs are hundreds of times less harmful to the ozone layer; therefore, they are still tolerated. A few HCFCs with two carbon atoms are still unknown; therefore, reliable methods for predicting their thermodynamic properties are relevant to the present needs of society.

Note Added in Proof. A further validation for the present NN models follows. For two HCFCs yet to be synthesized, predicted normal boiling points are for FCH_2–$CHCl_2$, 76.3°

(actual model 1) and 77.2° (actual model 2) and for FCH_2–CCl_3, 87.1° (actual model 1) and 85.0° (actual model 2). The predicted literature values (Woolf, A. A. Boiling Point Relations in the Halogenated Ethane Series. *J. Fluorine Chem.* **1990**, *50*, 89–99) are 74.4° and 88.8°, respectively, in good agreement with values obtained by the above NN models.

REFERENCES AND NOTES

(1) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists. An Introduction*; VCH Publishers: Weinheim, 1993.

(2) Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison-Wesley: New York, 1990.

(3) *Neural Network PC Tools. A Practical Guide*; Eberhart, R. C., Dobbins, R. W., Eds., Academic Press: San Diego, 1990.

(4) Zupan, J.; Gasteiger, J. Neural networks: A New Method for Solving Chemical Problems or Just a Passing phase? *Anal. Chim. Acta* **1991**, *248*, 1–30.

(5) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824–2836.

(6) Wikel, J. H.; Dow, E. R. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645–651.

(7) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 503–527.

(8) (a) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationship. *J. Med. Chem.* **1990**, *33*, 2583–2590. (b) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Structure-Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905–908.

(9) Livingstone, D. J.; Hesketh, G.; Clayworth, D. Novel Method for the Display of Multivariate Data Using Neural Network. *J. Mol. Graphics* **1991**, *9*, 115–118.

(10) Tetko, I. V.; Luik, A. I.; Poda, G. I. Applications of Neural Network in Structure-Activity Relationships of a Small Number of Molecules. *J. Med. Chem.* **1993**, *36*, 811–814.

(11) Darsey, J. A.; Noid, D. W.; Upadhyaya, B. R. Application of Neural Network Computing to the Solution for the Ground-State Eigenenergy of Two-Dimensional Harmonic Oscillators. *Chem. Phys. Lett.* **1991**, *177*, 189–194; Erratum: **1991**, *181*, 386.

(12) (a) Kvasnicka, V. An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **1991**, *6*, 63–76. (b) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of Recurrent Neural Networks in Chemistry. Prediction and Classification of ^{13}C NMR Chemical Shifts in a Series of Monosubstituted Benzenes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742–747.

(13) (a) Domine, D.; Devillers, J.; Chastrette, M.; Karcher, W. Estimating Pesticide Field Half-Lives from a Backpropagation Neural Network. *SAR and QSAR in Environmental Research*; **1993**, *1*, 211–219. (b) Hinton, G. E. How Neural Networks Learn from Experience. *Sci. Am.* **1992**, (Sept.), 145–151.

(14) Frank, J. E.; Friedman, J. H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993**, *35*, 109–135.

(15) (a) Wold, S.; Kettaneh-Wold, N.; Skagerrberg, B. Nonlinear PLS Modeling. *Chemometrics Intel. Lab. Syst.* **1989**, *7*, 53–65. (b) Wold, S. Nonlinear Partial Least Squares Modelling. II. Spline Inner Relation. *Chemometrics Intel. Lab. Syst.* **1992**, *14*, 71–84. (c) Wold, S. Discussion: PLS in Chemical Practice. *Technometrics* **1993**, *35*, 136–139.

(16) Balaban, A. T.; Joshi, N.; Kier, L. B.; Hall, L. H. Correlations between Chemical Structure and Normal Boiling Points of Halogenated Alkanes C_1–C_4. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 233–237.

(17) Balaban, A. T.; Kier, L. B.; Joshi, N. Correlations between Chemical Structure and Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals, and their Sulfur Analogues. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 237–244.

(18) Lohninger, H. Evaluation of Neural Network Based on Radial Basis Functions and Their Application to the Prediction of Boiling Points from Structural Parameters. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 736–744.

(19) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(20) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *80*, 399–404. Topological Indices Based on Topological Distance in Molecular Graphs. *Pure Appl. Chem.* **1983**, *55*, 199–206. Chemical Graphs. 48. Topological Index *J* for Heteroatom-Containing Molecules Taking into Account Periodicities of Element Properties. *Math. Chem. (MATCH)* **1986**, *21*, 115–122.

(21) Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the Structure-Function Relationship of Pharmacological and Toxicological Agents Using Information Theory. In *Proceedings of the Second International Conference on Mathematical Modeling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri at Rolla, 1979; pp 851–856.

(22) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Neighborhood Complexities and Symmetry of Chemical Graphs and Their Biological Applications. In *Mathematical Modeling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lipais, A. I., Rodin, E. Y., Eds.; Pergamon Press: 1984; p 745.

(23) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.