Development of Both Linear and Nonlinear Methods To Predict the Liquid Viscosity at 20 °C of Organic Compounds

Takahiro Suzuki,[†] Ralf-Uwe Ebert,[‡] and Gerrit Schüürmann*[,‡]

Research Laboratory of Resources Utilization, Tokyo Institute of Technology, 4259 Nagatsuta,
Midori-ku, Yokohama 226, Japan, and Department of Chemical Ecotoxicology, UFZ Centre for
Environmental Research, Permoserstrasse 15, D-04318 Leipzig, Germany

Received May 22, 1997[⊗]

Experimental values for the liquid viscosity (η) at 20 °C ranging from 0.164 mPa•s (*trans*-2-pentene) to 1490 mPa•s (glycerol) have been collected from literature for 361 organic compounds containing C, H, N, O, S, and all halogens. Multiple linear regression (MLR) and two-layer neural network (NN) modeling (one hidden layer) with back-propagation have been applied to derive prediction methods for log η using nine descriptors as input. The analysis includes different partitionings of the data set into training and prediction sets and different numbers of hidden-layer neurons of the neural networks. For the linear and nonlinear models derived from a training set of 237 compounds, squared correlation coefficients of 0.92 and 0.93 as well as root-mean-square errors of 0.17 and 0.16 log units were achieved for a prediction set of 124 compounds, reflecting a reasonable accuracy for a wide range of chemical structures and viscosity values. However, only the NN model was capable of successfully treating glycerol with the maximum viscosity value, which was not possible with the MLR approach and with any other existing estimation scheme.

INTRODUCTION

The viscosity of liquids is one of the key transport properties that is required in many scientific studies and engineering applications.[1] It has an important bearing on many problems relating to the transfer or movement of bulk quantities of the liquid. Consequently, viscosity data are becoming increasingly important in studies of the environmental behavior of organic compounds[2] and the quantitative structure−activity relationships (QSARs) of drugs.[3] In addition, chemical reactions in solution depend also on the viscosity of the solvent as described in Kramers' theory.[4]

For physicochemical properties like viscosity, most of the traditional estimation procedures have been based on one of the following three approaches: (1) equations derived from theoretical relationships, usually containing empirical parameters that have to be fitted; (2) additive−constitutive schemes based on atomic groups or bonds within molecules; and (3) linear or multilinear regression equations derived from the correlation of the property of interest with some other properties. With the advent of computers, various multivariate statistical tools, such as multiple linear regression, cluster analysis, principal component analysis, and partial least-squares regression, have been developed and applied to the study of quantitative structure−property relationships (QSPRs).[5−9] The QSPR philosophy assumes that the variation of behavior of organic compounds, as expressed by any measured physical or chemical properties, can be correlated with changes in molecular features of the compounds termed descriptors. While the traditional approach often needs some intuitive vision to derive the relevant mathematical relationship, QSPR methods are based on statistically determined linear or nonlinear functional forms that relate the property of interest with descriptors.

Recently, neural networks (NNs) have gained a great deal of interest in the field of QSPR.[10−15] NNs have an inherent ability to provide nonlinear and cross-product terms for QSPR modeling. In our previous paper, a predictive method for liquid viscosities of organic compounds based on the QSPR techniques using both multiple linear regression and partial least-squares regression was reported.[16] The purpose of this study is to extend the multilinear model by inclusion of an additional 124 compounds and to develop an alternative approach for predicting liquid viscosity by applying NN techniques. The prediction capabilities of both the linear and nonlinear approaches are tested explicitly by application of the models to subsets of compounds excluded from the training, and the discussion includes the dependence of the model performances on the degree of structural similarity between the training and prediction sets.

MATERIALS AND METHODS

Data Sets. Experimental liquid viscosities (η) at 20 °C of 237 diverse organic compounds containing C, H, O, N, S, and halogen atoms were taken from the previous work.[16] In this set, the range of experimental η values is 0.197−1490 mPa•s. An additional 124 compounds with experimental liquid viscosity values ranging from 0.164 to 130.3 mPa•s were collected from literature[1,17−20] and are listed in Table 1. For the first part of the modeling analyses, these latter 124 compounds served as the prediction set, and all compounds of the previous study[16] were used as the training set.

A second partitioning of the total set of 361 compounds into 237 training and 124 prediction compounds was generated under the guidance that the structural variety of both subsets is similar with regard to the relative portions of the

[*] To whom all correspondence should be addressed.

[†] Tokyo Institute of Technology.

[‡] UFZ Centre for Environmental Research.

[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1997.

S0095-2338(97)00446-0 CCC: $14.00 © 1997 American Chemical Society

Table 1. Data Set of 124 Additional Compounds (cf. Text) with Experimental Values for log η at 20 °C Taken from Literature

no.	compd name	CAS no.	log η_{exp}	ref
1	*cis*-2-pentene	627-20-3	−0.695	1
2	*trans*-2-pentene	646-04-8	−0.785	1
3	2-methyl-2-butene	513-35-9	−0.686	1
4	1,5-hexadiene	592-42-7	−0.561	18
5	*cis*-2-hexene	7688-21-3	−0.564	1
6	*trans*-2-hexene	4050-45-7	−0.564	1
7	1-undecene	821-95-4	0.013	20
8	1-dodecene	112-41-4	0.114	20
9	1-tridecene	2437-56-1	0.212	20
10	1-tetradecene	1120-39-1	0.301	20
11	1-pentadecene	13360-61-7	0.391	20
12	1-hexadecene	629-73-2	0.476	20
13	1-heptadecene	6765-39-5	0.556	20
14	1-octadecene	112-88-9	0.634	20
15	ethylcyclopentane	1640-89-7	−0.248	20
16	propylcyclopentane	2040-96-2	−0.167	20
17	butylcyclopentane	2040-95-1	−0.052	20
18	propylcyclohexane	1678-92-8	0.001	20
19	*trans*-1,3,5-trimethylcyclohexane	1839-63-0	−0.146	1
20	butylcyclohexane	1678-93-9	0.117	20
21	*n*-amylcyclopentane	3741-00-2	0.061	20
22	*n*-hexylcyclopentane	4457-00-5	0.274	20
23	*n*-amylcyclohexane	4292-92-6	0.235	20
24	*n*-heptylcyclopentane	5617-42-5	0.373	20
25	*n*-hexylcyclohexane	4292-75-5	0.344	20
26	*n*-octylcyclopentane	1795-20-6	0.464	20
27	*n*-heptylcyclohexane	5617-41-4	0.447	20
28	*n*-nonylcyclopentane	2882-98-6	0.550	20
29	*n*-octylcyclohexane	1795-15-9	0.544	20
30	*n*-decylcyclopentane	1795-21-7	0.550	20
31	*n*-nonylcyclohexane	2883-02-5	0.634	20
32	*n*-undecylcyclopentane	6785-23-5	0.631	20
33	*n*-decylcyclohexane	1795-16-0	0.719	20
34	*n*-dodecylcyclopentane	5634-30-0	0.708	20
35	*n*-undecylcyclohexane	54105-66-7	0.801	20
36	*n*-tridecylcyclopentane	6006-34-4	0.781	20
37	*n*-dodecylcyclohexane	1795-17-1	0.876	20
38	*n*-tetradecylcyclopentane	1795-22-8	0.852	20
39	*n*-tridecylcyclohexane	6006-33-3	0.949	20
40	*n*-pentadecylcyclopentane	4669-01-6	0.919	20
41	*n*-hexadecylcyclopentane	6812-39-1	0.982	20
42	α-methylstyrene	98-83-9	−0.099	1
43	1-methyl-4-ethylbenzene	622-96-8	−0.160	1
44	amylbenzene	538-68-1	0.124	20
45	1-phenylhexane	1077-16-3	0.223	20
46	1-phenylheptane	1078-71-3	0.316	20
47	1-phenyloctane	2189-60-8	0.408	20
48	1-phenylnonane	1081-77-2	0.496	20
49	1-phenyldecane	104-72-3	0.579	20
50	1-phenylundecane	6742-54-7	0.662	20
51	1-phenyldodecane	123-01-3	0.736	20
52	1-phenyltridecane	123-02-4	0.812	20
53	1-phenyltetradecane	1459-10-5	0.884	20
54	1-phenylpentadecane	2131-18-2	0.954	20
55	2-propyn-1-ol	107-19-7	0.225	18
56	allyl alcohol	107-18-6	0.134	17
57	2-methyl-1-butanol	34713-94-5	0.740	18
58	3-ethyl-3-pentanol	597-49-9	0.829	17
59	2-ethyl-1-hexanol	104-76-7	0.991	18
60	eugenol	97-53-0	0.965	17
61	1,3-butanediol	107-88-0	2.115	18
62	2-methyl-2,4-pentanediol	107-41-5	1.536	18
63	tripropylene glycol	1638-16-0	1.749	19
64	tetrahydropyran	142-68-7	−0.083	18
65	butyl vinyl ether	111-34-2	−0.301	18
66	dihexyl ether	112-58-3	0.232	1
67	dibenzyl ether	103-50-4	0.727	17
68	isobutyraldehyde	78-84-2	−0.246	1
69	2,6-dimethyl-4-heptanone	108-83-8	0.013	18
70	acrylic acid	79-10-7	0.114	18
71	pentanoic acid	109-52-4	0.350	20
72	2-methylbutyric acid	600-07-7	0.382	18
73	hexanoic acid	142-62-1	0.505	20
74	2-ethylbutyric acid	88-09-5	0.519	18
75	heptanoic acid	111-14-8	0.639	20
76	octanoic acid	124-07-2	0.766	18
77	2-ethylhexanoic acid	149-57-5	0.886	18
78	nonanoic acid	112-05-0	0.920	20
79	oleic acid	2027-47-6	1.589	19
80	vinyl formate	692-45-5	−0.444	1
81	methylmethacrylate	80-62-6	−0.199	18
82	ethyl acrylate	140-88-5	−0.210	18
83	isopropyl acetate	108-21-4	−0.245	18
84	methyl pentanoate	624-24-8	−0.147	18
85	2-methylbutyl acetate	624-41-9	−0.059	18
86	isoamyl acetate	123-92-2	−0.059	18
87	propyl butyrate	105-66-8	−0.080	18
88	propyl isobutyrate	644-49-5	−0.080	18
89	ethylpentanoate	539-82-2	−0.072	18
90	2-ethylhexyl acetate	103-09-3	0.176	18
91	butyl benzoate	136-60-7	0.493	1
92	dimethyl maleate	624-48-6	0.549	18
93	diethyl maleate	141-05-9	0.553	18
94	dibutyl maleate	105-76-0	0.751	19
95	diisobutyl *o*-phthalate	84-69-5	1.477	18
96	butyl decyl *o*-phthalate	89-19-0	1.740	18
97	4-methylpentanenitrile	542-54-1	−0.009	18
98	isopropylamine	75-31-0	−0.419	1
99	pentylamine	110-58-7	0.008	18
100	2-methylpyridine	109-06-8	−0.094	18
101	4-methylpyridine	108-89-4	−0.045	1
102	4-*tert*-butylpyridine	3978-81-2	0.175	18
103	*N*-butylaniline	1126-78-9	0.536	1
104	ethyl methyl sulfide	624-89-5	−0.428	18
105	tetrahydrothiophene	110-01-0	0.018	18
106	1,1-dichloroethylene	75-35-4	−0.446	18
107	*trans*-1,2-dichloroethylene	156-60-5	−0.394	18
108	1-bromopropane	106-94-5	−0.281	17
109	bromochlorobutane	74-97-5	−0.174	18
110	1,2,3-trichloropropane	96-18-4	0.406	1
111	1,1,2-trichlorotrifluoroethane	76-13-1	−0.148	18
112	2-methoxyethanol	109-86-4	0.236	18
113	2-mercaptoethanol	60-24-2	0.531	18
114	2-ethoxyethanol	110-80-5	0.312	19
115	methyl cyanoacetate	105-34-0	0.446	19
116	methyl acetoacetate	105-45-3	0.231	18
117	tetrahydropyran-2-methanol	100-72-1	1.041	18
118	2-hydroxybenzaldehyde	90-02-8	0.462	18
119	2,2′-thiodiethanol	111-48-8	1.814	18
120	2,2-dimethyl-1,3-dioxolane-4-methanol	100-79-8	1.041	18
121	bis(2-methoxyethyl) ether	111-96-6	0.299	18
122	*o*-phenetidine	94-70-2	0.784	17
123	*p*-phenetidine	156-43-4	1.111	17
124	1,2-bis(methoxyethoxy)ethane	112-49-2	0.575	18

major compound classes. For this partitioning, the software system *ChemProp*[21] was used; recent applications of ChemProp include the development of fragment-based estimation methods for water solubility[22] and vapor pressure[15] as well as an approach to improve existing schemes for calculation of octanol/water partition coefficient and water solubility through consideration of structural similarity.[23]

Within ChemProp, the generation of an optimized partitioning into training and prediction compounds contains two basic steps. First, the total compound set is sorted by predefined major compound classes and subclasses. From this sorted list, a predefined number of compounds forming the prediction list is selected randomly, but with two constraints: The two compounds with maximum and minimum target values are retained in the training set to ensure a proper range scaling (cf. eq 1), and chemical classes or subclasses with only two or one compound are also forced to be in the training set. It should be stressed that this partitioning is driven solely by the structural variety of the data set under analysis and does not include any knowledge about the subsequent modeling results.

Descriptors. The following nine descriptors were used for both multiple linear regression and neural network modeling: (1) molar refraction at 20 °C (MR, 10^{-6} $m^3 \cdot mol^{-1}$); (2) critical temperature (T_c, K); (3) absolute value of molar magnetic susceptibility (χ_m, 10^{-12} $m^3 \cdot mol^{-1}$); (4) cohesive energy (energy of vaporization) at 298 K (E_{coh}, $kJ \cdot mol^{-1}$); (5) indicator variable for alcohols/phenols (I_{OH}); (6) indicator variable for nitriles (I_{CN}); (7) indicator variable for amines (I_{amine}); (8) indicator variable for amides (I_{amide}); (9) indicator variable for aliphatic ring structures including O-, N-, and S-containing heterocycles (I_{ring}). The indicator variables are assigned values of 1 and 0 for the presence or absence of the relevant functional group except for polyols, where I_{OH} was set to 1.5 and 2 for dihydroxy and trihydroxy alcohols, respectively. In our previous study,[16] these nine parameters had been identified to be highly significant parameters for predicting liquid viscosity of organic compounds of the given range.

Multilinear Regression. Both the previous and current training and prediction sets were subjected to multilinear regression (MLR) of log η on the above-mentioned nine molecular descriptors, using the software package ChemProp.[21]

Neural Network Calculations. Two-layer neural networks (NNs) with nine input units plus a bias, a varying number of hidden-layer neurons (between 2 and 4 plus one bias), and one output neuron representing log η were optimized using the fast adaptive back-propagation algorithm[24] as implemented in ChemProp. For a more detailed description of the theory of back-propagation NNs and a number of practical applications, the reader is referred to the literature.[24,25] As with MLR, the training and prediction capability of NN models for log η was assessed using two different data set partitionings. The number of hidden-layer neurons was kept variable in the range mentioned to test its influence on the predictive quality of the NN model.[26]

All descriptor data x were transformed to values x' between 0.05 and 0.95 using

$$x' = 0.9\frac{x - x_{min}}{x_{max} - x_{min}} + 0.05 \tag{1}$$

and the same range-scaling formula was applied to the experimental log η values to yield proper target values y for the NN output. Adjustment of the weights during the training phase was performed after each individual compound. Following previous findings about the impact of the initial weights on the final NN model,[14] all NN calculations were performed with three different starting configurations, and the network output is calculated as the average of the output values of these three individual models.

A problem associated with the predictive capability of NN models is the question, how many iteration steps should be taken for the training phase? Convergence of the model error during training may include substantial overtraining, which is only seen with truly predictive applications of the network.[14] In order to avoid overtraining, the prediction performance in terms of the relative global error,

$$\%_{global} = \frac{SE}{y_{max} - y_{min}} \times 100 \tag{2}$$

was monitored for both the training and prediction set during the training phase. Here, SE is the standard error,

$$SE = \left[\frac{1}{3n - 1}\sum_{c=1}^{3}\sum_{i=1}^{n}(y_i - y_{ic}^{cal})^2\right]^{1/2} \tag{3}$$

which includes averaging over three different starting configurations for the weights as mentioned above; in eq 3, y_{ic}^{cal} denotes the calculated value for a given starting configuration (c) and compound (i). The final network was selected from a maximum of 400 000 iteration cycles such that, at the optimal training step, the sum of relative global errors for the training and prediction sets is minimal. All back-propagation NN runs were performed with a learning rate of 0.10 and a momentum term of 0.10.

Statistical Parameters. The statistical quality of the MLR and NN modeling results for both training and prediction sets was evaluated using the following parameters: Squared correlation coefficient r^2,

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{fit})^2}{\sum_{i=1}^{n}(y_i - y_0)^2} \tag{4}$$

root-mean-square error RMSE,

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{fit})^2\right]^{1/2} \tag{5}$$

average absolute error AAE,

$$AAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i^{fit}| \tag{6}$$

and bias,

$$bias = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^{fit}) \tag{7}$$

In these formulas, y_i represents the experimental target value (log η) for the ith compound, y_0 denotes the associated mean (with n being 237 and 124 for the training and prediction sets, respectively), and y_i^{fit} represents the calculated target value using the NN or MLR model. In order to make the comparison between training and prediction quality as simple as possible, r^2 and RMSE do not contain any correction for the number of degrees of freedom.

RESULTS AND DISCUSSION

Compound Class Characteristics of the Data Set. The entire data set contains 119 hydrocarbons, 44 halogenated hydrocarbons, and 143 oxygen-containing, 44 nitrogen-containing, and 11 sulfur-containing compounds with different functional groups.

A more detailed analysis of the chemical class distribution in the training and prediction sets is given in Table 2. As

Table 2. Compound Class Characteristics of the Different Training and Prediction Sets[a]

	previous partitioning[b]				current partitioning[c]			
	training		prediction		training		prediction	
compd class								
hydrocarbons	65	24.4%	54	43.5%	78	32.9%	41	33.0%
nonaromatic	49	20.7%	41	33.1%	60	23.3%	30	24.2%
aromatic	16	6.7%	13	10.5%	18	7.6%	11	8.9%
halogenated hydrocarbons	38	16.0%	6	4.8%	30	12.7%	14	11.3%
nonaromatic	28	11.8%	6	4.8%	24	10.1%	10	8.1%
aromatic	10	4.2%	0	0%	6	2.5%	4	3.2%
alcohols/phenols	22	9.3%	7	5.6%	19	8.0%	10	8.1%
aldehydes/ketones	16	6.7%	2	1.6%	12	5.1%	6	4.8%
carboxylic acids/esters/anhydrides	32	13.5%	27	21.8%	38	16.0%	21	16.9%
ethers/furanes	16	6.7%	6	4.8%	14	5.9%	8	6.5%
mixed oxygen compounds	3	1.3%	8	6.5%	9	4.0%	2	1.6%
halogenated compounds with oxygen	4	1.7%	0	0%	3	1.3%	1	0.8%
amines/anilines/azols/azines	18	7.6%	6	4.8%	15	6.3%	9	7.3%
nitriles/nitro compounds	11	4.6%	1	0.8%	7	3.0%	5	4.0%
amides/(mixed N + O)/(N + halogen)	5	2.1%	3	2.4%	5	2.1%	3	2.4%
sulfur compounds	7	3.0%	4	3.2%	7	3.0%	4	3.2%

[a] The total set of 361 compounds was subdivided in two different ways in training and predicting sets containing 237 and 124 compounds, respectively (cf. text). The column entries give the absolute and relative number of compounds with certain structural features in each of the four subsets, where the relative numbers represent the percentages of compounds in the data sets. [b] The training set of the previous partitioning contains all 237 compounds of the previous study,[16] and the prediction set contains all 124 compounds listed in Table 1. [c] The prediction set of the compound-class oriented partitioning contains the following compounds from the previous training and prediction sets, identified by their numbers according to those given in the previous study[16] and in Table 1, respectively. Compounds from previous training set: 1, 3, 6, 11, 17, 23, 29, 31, 35, 36, 38, 41, 43, 44, 46, 50, 53, 58, 60, 61, 63, 65, 67, 70, 75, 77, 78, 80, 86, 87, 94, 97, 99, 101, 105, 109, 112, 114, 117, 119, 128, 129, 134, 135, 139, 143, 148, 150, 156, 159, 167, 172−175, 178, 182, 187, 189, 190−192, 194, 196, 203, 205−209, 211, 216, 222, 225, 228, 230, 234, 236. Compounds from previous prediction set: 1, 3, 6, 7, 10, 11, 15, 16, 21, 22, 24, 26, 29, 31, 33, 34, 37, 39, 42, 46, 49, 52, 55, 61, 67, 72, 75, 78, 80, 83, 87, 88, 90, 93, 99, 100, 107, 115−117, 119, 124. Correspondingly, the training set of this data set partitioning contains the remainder of 237 compounds.

Table 3. Statistics of Multilinear Regression (MLR) Models and Neural Network (NN) Models with Three Hidden-Layer Neurons for Calculating Liquid Viscosity at 20 °C with the Previous and Current Partitioning into Training and Prediction Sets[a]

	MLR $n_{\text{descr}} = 9$, $n_{\text{param}} = 10$		NN $n_{\text{descr}} = 9$, $n_{\text{param}} = 34$	
	training	prediction	training	prediction
		Previous Partitioning		
n	237	124	237	124
r^2	0.922	0.867	0.955	0.868
RMSE	0.158	0.201	0.120	0.201
AAE	0.102	0.152	0.084	0.133
bias	0	0.045	0.000	0.027
error range	−0.91 to +0.58	−0.49 to +0.82	−0.47 to +0.36	−0.54 to +1.00
		Current Partitioning		
n	237	124	237	124
r^2	0.916	0.919	0.958	0.926
RMSE	0.167	0.168	0.118	0.161
AAE	0.109	0.107	0.084	0.105
bias	0	−0.016	0.000	−0.008
error range	−1.08 to +0.70	−0.89 to +0.34	−0.47 to +0.43	−0.86 to +0.35

[a] The previous and current partitionings of the total of 361 compounds into training and prediction sets are described in the text and in Table 1. Abbreviations: n_{descr} = no. of descriptors, n_{param} = no. of model parameters, n = no. of compounds, r^2 = squared correlation coefficient without consideration of degrees of freedom (eq 4), RMSE = root-mean-square error (eq 5), and AAE = average absolute error (eq 6). The bias was calculated according to eq 7, and the error range is defined by the greatest underestimations (negative values) and overestimations (positive values) of log η.

can be seen from the table, the relative portion of hydrocarbons, carboxylic acids and esters, and mixed oxygen compounds was considerably greater in the previous prediction set than in the associated training set, and the reverse was true for halogenated hydrocarbons, aldehydes, and ketones as well as for nitriles and nitro compounds. The new partitioning yields a clearly more balanced distribution of the chemical classes among the training and prediction sets, which were generated using ChemProp[21] and is given in the right part of Table 2. As shown below, this new partitioning leads to considerably improved performance of linear and nonlinear models in predicting liquid viscosity.

MLR Models. Multilinear regression of log η against the nine descriptors yields the following equation for the current training set of 237 compounds:

$$\log \eta = -0.0353\text{MR} + 0.00346T_{\text{c}} + 0.00083\chi_{\text{M}} + 0.0158E_{\text{coh}} + 0.452I_{\text{OH}} - 0.181I_{\text{CN}} + 0.116I_{\text{amine}} + 0.364I_{\text{amide}} + 0.0837I_{\text{ring}} - 2.438 \quad (8)$$

The respective statistics are summarized in Table 3 and compared with the MLR performance on the basis of the previous partitioning. The somewhat better fit with the previous training set is opposed to a significantly improved

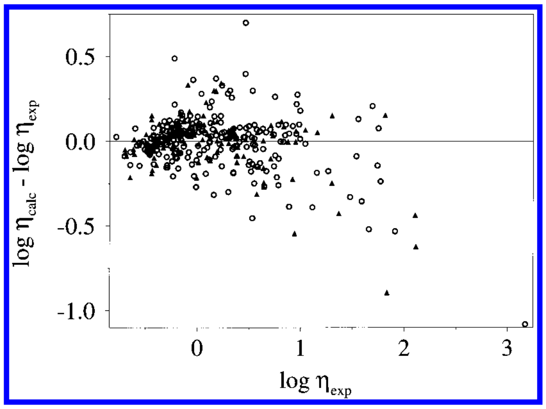

Figure 1. Calculation errors vs experimental values of log η for the training set (circles) and prediction set (triangles) of the current partitioning (cf. Table 2), using the MLR model of eq 8.

Table 4. Weights of the NN Model with Three Hidden-Layer Neurons To Predict Liquid Viscosity at 20 °C of Organic Compounds[a]

	hidden-layer neuron			
neuron	1	2	3	bias
	First Starting Configuration			
input layer				
1	1.7958	3.6902	7.2307	na
2	−2.2815	0.1855	1.5677	na
3	−4.8182	0.8491	−4.0831	na
4	3.2763	−6.1663	−2.0297	na
5	0.4564	−2.4224	−11.257	na
6	0.7679	−0.9553	0.9768	na
7	2.2567	1.3546	−18.952	na
8	−5.6587	1.2489	−0.5653	na
9	−0.2429	0.0239	2.0920	na
bias	−0.0219	−0.0847	9.4030	na
output layer				
1	−5.0601	−4.1568	−3.6290	5.469
	Second Starting Configuration			
input layer				
1	7.6318	3.6100	1.7751	na
2	1.5346	0.1763	−2.2849	na
3	−4.5589	0.7968	−4.7886	na
4	−1.9428	−6.0018	3.2836	na
5	−11.062	−2.4062	0.5031	na
6	1.1050	−0.9512	0.7828	na
7	−18.890	1.3645	2.2652	na
8	−0.5813	1.2072	−5.6685	na
9	2.1580	0.0252	−0.2458	na
bias	9.2531	−0.0949	−0.0341	na
output layer				
1	−3.6776	−4.2683	−5.0607	5.5466
	Third Starting Configuration			
input layer				
1	6.9728	3.6019	1.7801	na
2	1.6019	0.0620	−2.2838	na
3	−3.8998	0.6108	−4.8514	na
4	−2.1650	−5.7621	3.3474	na
5	−11.390	−2.5662	0.6534	na
6	1.0824	−0.9568	0.8355	na
7	−19.509	1.2833	2.4350	na
8	−3.1917	1.0833	−5.5610	na
9	2.2747	0.0143	−0.2464	na
bias	9.6674	0.0060	−0.1442	na
output layer				
1	−3.7363	−4.3338	−4.9909	5.5798

[a] The final NN model consists of three individual submodels according to three different starting configurations, and the NN model output is calculated as the average of the output values of these three individual models.[14] [b] na denotes "not applicable".

prediction capability derived from the current partitioning, which is seen by the greater r^2 (0.919 vs 0.867) as well as by smaller values for RMSE (0.168 vs 0.201), AAE (0.107 vs 0.152), and the bias (−0.016 vs +0.045). In particular, the ratio of prediction RMSE over training RMSE is 1.27 for the previous partitioning and 1.01 for the compound-class oriented partitioning. Overall, the latter subdivision into training and prediction sets yields a clearly better MLR model. This result reveals that a judicious partitioning means of the data set is crucial in the development process of statistically sound models.

The data distribution of calculation errors vs experimental values is plotted in Figure 1. The greatest overestimations of log η are observed for the training compounds methanol (0.489) and 2-hydroxybenzaldehyde (0.701), and the greatest underestimations for the training compounds 1-isopropyl-4-methylbenzene (−0.453), 1,2-propanediol (−0.520), bis(2-ethylhexyl)-*o*-phthalate (−0.532), and glycerol (−1.081), as well as for the prediction compounds ethylcinnamate (−0.545), 1,5-pentanediol (−0.437), 1,3-butanediol (−0.622), and cyclohexanol (−0.894). This list of outliers suggests that compounds with several OH groups may need a more elaborated parametrization for the effect of the hydrogen bond, which however would require inclusion of an additional set of compounds.

The RMSE values of eq 8 correspond to an uncertainty factor around 1.5 for predicting liquid viscosity through application of this MLR model, provided that the chemical functionalities of the compounds are covered in the present data set. To our best knowledge, the chemical domain and associated viscosity range covered by eq 8 is greater than with any other currently available additive scheme to calculate liquid viscosity at 20 °C.

NN Models. The statistical results of the NN modeling are also listed in Table 3 for comparison with the MLR models, and the relevant weights of the NN model derived from the current data set partitioning are listed in Table 4. The final network architecture is (9 + 1):(3 + 1):1 and thus contains nine input units plus a bias, three hidden-layer neurons plus a bias, and one output layer neuron. With this architecture, the NN model contains a total number of 34 adjustable parameters. Thus, the number of training compounds (237) is seven times greater than the number of model parameters.

The development of the relative global training and prediction errors (eq 2) with increasing numbers of iteration cycles is shown in Figure 2 for the current data set partitioning. In this case, the minimum of the error sum is achieved after 59 000 training steps, and all associated NN model results presented below (Tables 3 and 4 as well as Figure 3) refer to this training status.

As can be seen from Table 3, training of the NN model yields r^2 values around 0.96 and RMSE values of ca. 0.12 log units of η, indicating a significantly better fit than was achieved with MLR for both the previous and current training sets. On the other hand, the prediction performance is just comparable to the MLR results and only slightly better than the linear model for the new data set partitioning. As with MLR, the preferable compound-class subdivision of available data into a training and prediction set leads to a great improvement of the predictive power.

The calculation error is plotted against experimental log η in Figure 3. Comparison of Figures 1 and 3 shows a

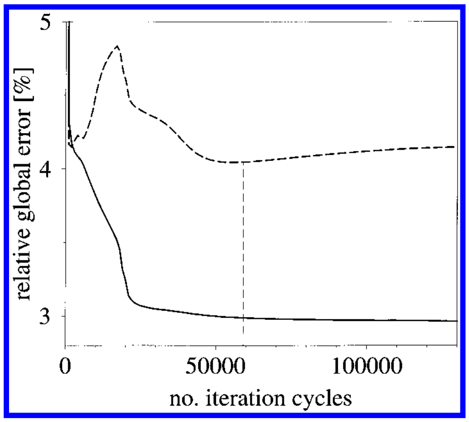

Figure 2. Relative global errors (eq 2) of the NN model with three hidden-layer neurons for the chemical-class oriented partitioning (cf. Table 2) into a training set (solid curve) and prediction set (broken curve) as a function of the number of iteration cycles (training steps). At 59 000 training steps as indicated by the dashed vertical line, the error sum is 2.99% (training) + 4.05% (prediction) = 7.04%.

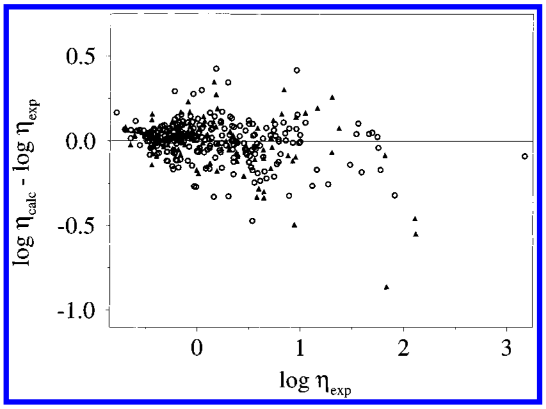

Figure 3. Calculation errors vs experimental values of log η for the training set (circles) and prediction set (triangles) of the current partitioning (cf. Table 2), using the NN model with three hidden-layer neurons as specified in Table 4.

significant improvement for some of the previous outliers. The case of glycerol having the highest reported viscosity value of 1490 mPa•s is particularly striking: With MLR this training compound showed the greatest calculation error of all 361 compounds, while the NN model yields a much better performance with a quite small calculation error of only −0.090 log units of η. On the other hand, a substantial underestimation of log η is again observed for the training compound 1-isopropyl-4-methylbenzene (−0.472), and the prediction set still contains four outliers with significant underestimations of log η: ethylcinnamate (−0.494), 1,5-pentanediol (−0.458), 1,3-butanediol (−0.548), and cyclohexanol (−0.861). This error pattern might reflect apparent deficiencies of the current NN model with multiple OH groups (see above), but the case of cyclohexanol might also indicate a possible problem with the experimental value.

Interestingly, the difference between the recognition and prediction power is much greater for the NN model than for MLR. From the viewpoint of the large difference between the number of model parameters of NN (34) and MLR (10), one could expect that alternative NN architectures with smaller numbers of model parameters would yield increased r^2 values for the prediction set. To our surprise, a corresponding analysis based on the current data set partitioning with only two hidden-layer neurons, that is a total of 23 model parameters, gave significantly inferior statistics for the prediction:

2 hidden-layer neurons, 23 adjustable parameters:

training $r^2 = 0.950$, RMSE = 0.128, AAE = 0.091, bias = 0.001

prediction $r^2 = 0.895$, RMSE = 0.191, AAE = 0.112, bias = −0.020

Comparison with Table 3 shows further that these results cannot compete with the prediction performance of the (still much simpler) MLR model.

The alternative NN model with four hidden-layer neurons yields the following results:

4 hidden-layer neurons, 45 adjustable parameters:

training $r^2 = 0.960$, RMSE = 0.115, AAE = 0.082, bias = 0.000

prediction $r^2 = 0.922$, RMSE = 0.164, AAE = 0.104, bias = −0.014

Both training and prediction performances are close to the results with three hidden-layer neurons and, in particular, better than with only two hidden-layer neurons. It shows that, for some reason, the NN architecture with 23 model parameters is inferior to both less and more complex models (being represented by MLR and the NN models with more hidden-layer neurons, respectively). Further studies may show whether an improved performance of this architecture could be possible through selection of some other network parameters (learning rate, momentum, optimization algorithm).

With regard to the application range and overall performance, both the linear model (eq 8) and the nonlinear model (Table 4) are superior to other currently available models[1,2] in predicting liquid viscosity at 20 °C. Furthermore, a parallel use of both models may help in identifying compounds for which predictions of η from the nine descriptors could be less reliable.

The viscosity of liquids generally decreases with the temperature, which can be approximately expressed by corresponding empirical relationships.[1] To obtain viscosities at temperatures different from 20 °C, the MLR approach would require another treatment,[16] while the NN approach could include the nonlinear temperature effect as an input parameter into the architecture.

CONCLUSIONS

The comparative analysis of MLR and NN model performances with different training and prediction sets demonstrates, that a compound-class oriented data set partitioning may be crucial in enabling derivation of statistically sound structure−property relationships. With the present data set of 361 compounds, recognition and prediction capabilities are almost identical for MLR but significantly different for the NN models. This suggests that there is still room for improvement of the nonlinear model using the same set of descriptors through inclusion of additional compounds of the same chemical domain. The presently derived models allow predictive applications with expected uncertainty factors for η of 1.5 (MLR) and 1.4 (NN), respectively, which is reasonable accuracy for the wide range of chemical structures with η values covering 4 orders of magnitude.

REFERENCES AND NOTES

(1) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987.

(2) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*; McGraw-Hill: New York, 1982.

(3) Dearden, J. C. Applications of Quantitative Structure-Property Relationships to Pharmaceutics. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 77−87.

(4) Billing, G. D.; Mikkelsen, K. V. *Introduction to Molecular Dynamics and Chemical Kinetics*; Wiley: New York, 1996.

(5) Aries, R. E.; Lidiard, D. P.; Spragg, R. A. Principal component analysis. *Chem. Br.* **1991**, 821−824.

(6) Wold, S. PLS for multivariate linear modelling. In *Chemometric methods in molecular design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195−218.

(7) Eriksson, L.; Johansson, E.; Wold, S. QSAR model validation. In *QSAR in Environmental-VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola, FL, in press.

(8) Balzuweit, G.; Welk, M.; Der, R.; Schüürmann, G. Nonlinear partial least-squares regression. In *Solving engineering problems with neural networks. Proceedings of the International Conference EANN'96*; Bulsari, A. B., Kallio, S., Tsaptsinos, D., Eds.; London 1996; pp 495−498.

(9) Schüürmann, G.; Segner, H.; Jung, K. Multivariate Mode-of-Action Analysis of Acute Toxicity of Phenols. *Aquat. Toxicol.* **1997**, *38*, 277−296.

(10) Bodor, N.; Harget, A.; Huang, M.-J. Neural Network Studies: 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.

(11) Egolf, L. M.; Jurs, P. C. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616−625.

(12) Sigman, M. E.; Rives, S. S. Prediction of Atomic Ionization Potentials I−III Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 617−620.

(13) Gakh, A. A.; Gakh, E. G.; Sumpter, B. G.; Noid, D. W. Neural Network−Graph Theory Approach to the Prediction of Physical Properties of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 832−839.

(14) Schüürmann, G.; Müller, E. Back-Propagation Neural-Networks-Recognition vs. Prediction Capability. *Environ. Toxicol. Chem.* **1994**, *13*, 743−747.

(15) Kühne, R.; Ebert, R.-U.; Schüürmann, G. Estimation of vapour pressures for hydrocarbons and halogenated hydrocarbons from chemical structure by a neural network. *Chemosphere* **1997**, *34*, 671−686.

(16) Suzuki, T.; Ohtaguchi, K.; Koide, K. Computer-Assisted Approach to Develop a New Prediction Method of Liquid Viscosity of Organic Compounds. *Comput. Chem. Eng.* **1996**, *20*, 161−173.

(17) Weast, R. C., Ed. *CRC Handbook of Chemistry and Physics*, 69th ed.; CRC: Boca Raton, FL, 1988−1989.

(18) Dean, J. A. *Handbook of Organic Chemistry*; McGraw-Hill: New York, 1987.

(19) Dean, J. A., Ed. *Lange's Handbook of Chemistry*, 13th ed.; McGraw-Hill: New York, 1985.

(20) Pachaiyappan, V.; Ibrahim, S. H.; Kuloor, N. R. Simple correlation for determining viscosity of organic liquids. *Chem. Eng.* **1967**, *74*, 193−196.

(21) Schüürmann, G.; Kühne, R.; Kleint, F.; Ebert, R.-U.; Rothenbacher, C.; Herth, P. A Software System for Automatic Chemical Property Estimation from Molecular Structure. In *QSAR in Environmental Sciences-VII*; Chen, F., Schüürmann, G., Eds.; SETAC Press: Pensacola FL, in press.

(22) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group contribution methods to estimate solubility of organic chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(23) Kühne, R.; Kleint, F.; Ebert, R.-U.; Schüürmann, G. Calculation of Compound Properties Using Experimental Data From Sufficiently Similar Chemicals. In *Software development in chemistry 10*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker (GDCh): Frankfurt, Germany, 1996; pp 125−134.

(24) Tollenaere, T. Fast adaptive backpropagation with good scaling properties. *Neural Networks* **1990**, *3*, 561−573.

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

(26) Livingstone, D. L.; Salt, D. W. Regression Analysis for QSAR Using Neural Networks. *Bioorg. Med. Chem. Lett.* **1993**, *3*, 645−651.

CI9704468