ELSEVIER

Journal of Chromatography A, 950 (2002) 183–194

JOURNAL OF CHROMATOGRAPHY A

www.elsevier.com/locate/chroma

Prediction of relative response factors for flame ionization and photoionization detection using self-training artificial neural networks

M. Jalali-Heravi*, Z. Garkani-Nejad

Department of Chemistry, Sharif University of Technology, P.O. Box 11365-9516, Tehran, Iran

Received 14 August 2001; received in revised form 3 January 2002; accepted 10 January 2002

Abstract

The relative response factors (RRFs) of a flame ionization detection (FID) system and two pulsed discharge photoionization detection (PID) systems with different discharge gases are predicted for a set of organic compounds containing various functional groups. As a first step, numerical descriptors were calculated based on the molecular structures of compounds. Then, multiple linear regression (MLR) was employed to find informative subsets of descriptors that can predict the RRFs of these compounds. The selected MLR model for the FID system includes seven descriptors and two selected MLR models for the PID systems with argon- and krypton-doped helium as the discharge gases, respectively, include six and five descriptors. The descriptors appearing in the MLR models were considered as inputs for the self-training artificial neural networks (STANNs). A 7-7-1 STANN was generated for prediction of RRFs of the FID system, and two STANNs with the topologies of 6-7-1 and 5-6-1 were generated for the two PID systems. Comparison of the results indicates the superiority of neural networks over that of the MLR method. This is due to the nonlinear behaviors of relative response factors for all type of detectors studied in this work. © 2002 Published by Elsevier Science B.V.

Keywords: Neural networks, artificial, self-training; Response factors; Flame ionization detection; Photoionization detection; Detection, GC; Regression analysis; Molecular descriptors

1. Introduction

The development of sensitive and selective detectors has played a major role in the establishment of chromatography as an unrivaled analytical tool. The retention time can be used for identification of compounds, but it is well accepted that more than one compound can have a similar retention time. However, the different detector responses can be used for peak identification of compounds with the same retention time. On the other hand, the response factor (RF) is essentially a correction factor, which measures the response of a given compound to the detecting device.

Since numerous compounds are unavailable as standards, the development of a theoretical method for estimating response factor seems to be useful. The first work on the prediction of response factors of substituted benzenes and pyridines using a multivariate statistical partial least-squares (PLS) treat-

*Corresponding author. Tel.: +98-21-6005-718; fax: +98-21-6012-983.

E-mail address: jalali@sina.sharif.ac.ir (M. Jalali-Heravi).

0021-9673/02/$ – see front matter
PII: S0021-9673(02)00054-7

ment was published by Katritzky and Gordeeva [1]. Also, Katritzky and coworkers applied the multiple linear and nonlinear regression methods to predict the retention time and response factors of different organic compounds [2,3]. Jalali-Heravi and Fatemi have used artificial neural networks (ANNs) for predicting flame ionization detection (FID) and thermal conductivity detection (TCD) response factors for different series of organic molecules [4,5].

The use of the pulsed discharge source in a photoionization detector has been described in the literature [6,7]. The photon emission arises from the discharge and is dependent upon the composition of the make-up gas passing through the discharge. When pure helium is used as the make-up gas the emission arises from the following transition (Eq. (1)) [8]:

$$\mathrm{He}_2\,(A^1\Sigma_u) \rightarrow 2\mathrm{He}\,(1\,{}^1S_0) \qquad (1)$$

If the make-up gas passing through the discharge region is doped with argon or krypton, the emission spectra consists principally of the resonance lines from the argon or krypton. The photoionization detection (PID) response is proportional to the number of electrons in the molecule with energies less than the photoionization energies coming from the discharge. Since, the emission lines of Ar_2 and Kr_2 are different, the response of compounds to the Ar-PID and Kr-PID systems are not the same.

The main aim of the present work was the development of a quantitative–structure property relationship (QSPR) using for the first time a self-training artificial neural network (STANN) for predicting relative response factors (RRFs) for different detection system. In this study, RRFs for FID and PID systems with argon- and krypton-doped helium as discharge gases (Ar-PID, Kr-PID) were predicted for a diverse set of organic compounds.

2. Methods

Artificial neural networks (ANNs) are mathematical systems that simulate biological neural networks [9–11]. They consist of processing elements (nodes, neurons) organized in layers. Back-propagation neural networks (BNNs) are most often used in analytical applications. The back-propagation network receives a set of inputs, which are multiplied by each node's weights. These products are summed for each node and then a nonlinear transfer function is applied. The goal of training the network is to change the weights between the layers in a direction that minimizes the output errors. The changes in the values of the weights can be obtained using Eq. (2):

$$\Delta W_{ij}(n) = \eta\delta_i O_j + \alpha\Delta W_{ij}(n-1) \qquad (2)$$

where ΔW_{ij} is the change in the weight factor for each network node, δ_i is the actual error of node i, and O_j is output of node j. The coefficients η and α are the learning rate and the momentum factor, respectively.

A self-training artificial neural network (STANN) [12] is a new method for updating the node's weights and training of the networks in parallel fashion. In the STANN, the important aspect is a network, which trains another networks. The architecture of a STANN is shown in Fig. 1. The structure of network

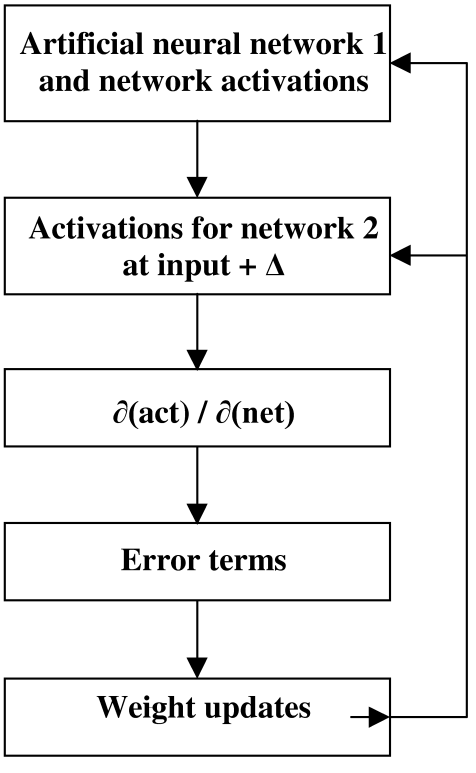

Fig. 1. The architecture of a STANN.

Table 1
Compounds studied in this work

No.	Compound	No.	Compound
Training set			
1	Dibromomethane	62	3,3-Dimethylpentane
2	$CHCl_2CH_2Cl$	63	1-Heptene
3	CCl_3CH_3	64	Toluene
4	1,3-Dibromopropane	65	2,4-Dimethylpentane
5	$CFCl_2CF_2Cl$	66	Diisopropyl ether
6	Ethyl disulfide	67	*p*-Xylene
7	3-Bromopentane	68	Trimethylacetonitrile
8	Idomethane	69	Ethyl benzene
9	2-Bromopentane	70	*o*-Xylene
10	1-Pentanethiol	71	Cumene
11	Ethyl iodide	72	*n*-Butylbenzene
12	1-Butanol	73	3-Hexyne
13	CH_2ClCH_2Cl	74	3-Ethyl-1-pentene
14	1-Bromobutane	75	Cyclohexene
15	1-Bromopentane	76	Butyraldehyde
16	2-Hexanone	77	*sec*.-Butylbenzene
17	Cyclopentylchloride	78	Methyl *tert*.-butyl ether
18	2-Methylheptane	79	Isopropanol
19	2-Nitropropane	80	Propionitrile
20	Butyl formate	81	1-Chloropropane
21	2-Ethylbutyraldehyde	82	Propionaldehyde
22	2,3,4-Trimethylpentane	83	2-Butanone
23	*cis*-CHCl=CHCl	84	Ethanol
24	Propyl sulfide	85	Bicylco[2,2,1]hepta-2,5-diene
25	Cycloheptane	86	Methylcyclopentane
26	Propyl acetate	87	Crotonaldehyde
27	1,3-Dichlorobenzene	88	Hexane
28	$CHCl_2CH_3$	89	2-Chloropropane
29	Ethylcyclohexane	90	α,α,α-Trifluorotoluene
30	Dipropyl ether	91	1-Hexyne
31	Isopropyl acetate	92	2-Methyl-1-pentene
32	2,3-Butanedione	93	1-Hexene
33	Nitromethane	94	2,2-Dimethylbutane
34	2-Methyl-1-propanol	95	*m*-Xylene
35	Cyclopropylcyanide	96	2-Methyl-2-propanol
36	Allylsulfide	97	Acetonitrile
37	1,2-Dichlorobenzene	98	Methacrylonitrile
38	1-Methylcyclohexene	99	Cyclopentane
39	4-Methylcyclohexene	100	Pentane
40	Methanol	101	2-Methyl-2-butene
41	Methyl propionate	102	1,2-Difluorobenzene
42	1-Ethylcyclopentene	103	Acrylonitrile
43	3-Pentanone	104	1-Pentene
44	1-Bromopropane	105	Hexafluorobenzene
45	*trans*-CHCl=CHCl	*Prediction set*	
46	Allyl acetate	106	CH_2Cl_2
47	Valeraldehyde	107	$CH_2ClCHClCH_3$
48	Ethyl acetate	108	Ethyl sulfide
49	3,3-Dimethyl-2-butanone	109	1,1-Dimethylcyclohexane
50	2,2,4-Trimethylpentane	110	3,3-Diethylpentane
51	3-Ethylpentane	111	Heptane
52	*trans*-2-Heptene	112	Butyronitrile
53	*sec*.-Butanol	113	1-Octyne
54	2-Pentanone	114	Propyl formate
55	3-Ethyl-2-pentene	115	1-Propanol
56	Acetaldehyde	116	Tetrahydrofuran
57	4-Bromo-*m*-xylene	117	Isobutyronitrile
58	1-Heptyne	118	2-Hexyne
59	2-Methyl-2-butanol	119	Propyl benzene
60	2-Bromo-*p*-xylene	120	Diethyl ether
61	2-Bromopropane	121	Acetone
		122	Cyclopentene

2 is similar to a back-propagation artificial neural network (BNN). However, during the training the normalized inputs are changed by some infinitesimal amount delta (Δ). In this regard, because the transfer function being utilized, a sigmoid, has a linear region around the value 0.5, it is desirable when adding the delta value to the normalized input to adjust the input towards the linear region. Thus, the positive delta value should be added to normalized inputs which are less than 0.5 and the negative delta values should be added to normalized inputs which are greater than 0.5. For the hidden layer a similar manner is used. The network 1 uses from weight updates produced by the training network 2. Thus, training of the artificial neural network 1 is not carried out with algorithmic code, but rather by a network training a network.

In a previous work we have compared the performance of the STANN with the conventional ANN in predicting the gas chromatographic relative retention times of a variety of organic compounds [13]. It was shown that using a STANN reduces the number of the adjustable parameters in the network and the optimization procedure was faster compared with the conventional ANN. The aim of this work was to examine this conclusion in predicting of the FID and PID relative response factors for various organic compounds, which the mechanisms of which show some nonlinear characteristics.

3. Experimental

3.1. Data set

FID, Ar-PID and Kr-PID RRFs were taken from Ref. [14]. This data set consists of 13 different classes of organic compounds containing numerous functional groups such as alcohols, ketones, aldehydes, esters, alkenes, alkynes, alkanes, halides, thiols, nitros, ethers, cyanides, and sulfides. These molecules were randomly divided into two groups: a training set and a prediction set (Table 1). The training and prediction sets consist of 105 and 17 compounds, respectively. The training set was used for the generation of models and the prediction set was used for the evaluation of the generated models. The prediction set consists almost all types of

Table 2
Specifications of the selected multiple linear regression models for different detectors

Detection method	Descriptor	Notation	Coefficient	Mean effect
FID	Boiling point	b.p.	−0.003 (±0.000)	−0.338
	Maximum bond order of C–X	MBOC	1.289 (±0.187)	2.655
	Path one connectivity index	$^{1}\chi_{p}$	−0.099 (±0.022)	−0.246
	polarizability	α	−0.085 (±0.030)	−0.601
	Square of polarizability	α^{2}	0.006 (±0.002)	0.351
	Relative number of C atoms	RENC	1.422 (±0.224)	0.468
	Maximum valency of H atoms	MVH	−0.614 (±0.153)	−0.602
	Constant		−1.075 (±0.246)	
Ar-PID	Molecular density	MD	−0.920 (±0.110)	−0.959
	Dipole moment	DIMO	−0.065 (±0.018)	−0.085
	Cluster three connectivity index	$^{3}\chi_{c}$	−0.117 (±0.039)	−0.033
	Heat of formation	ΔH	0.002 (±0.000)	−0.073
	Highest occupied molecular orbital	HOMO	0.085 (±0.024)	−0.887
	Distance between center of mass and center of charge	DMCH	−0.134 (±0.050)	−0.055
	Constant		2.833 (±0.278)	
Kr-PID	Molecular density	MD	−1.015 (±0.164)	−1.073
	Path four connectivity index	$^{4}\chi_{p}$	−0.193 (±0.066)	−0.071
	Highest occupied molecular orbital	HOMO	0.221 (±0.027)	−2.310
	Principal moment of inertia about x axis	MO_{x}	0.163 (±0.026)	0.076
	Relative weight of effective C atoms	REWC	0.663 (±0.153)	0.351
	Constant		3.303 (±0.294)	

molecules included in the training set and therefore, is a good representative of the training set.

3.2. Descriptor generation

A total of 77 separate molecular structure descriptors were calculated for each compound in the data set. These descriptors can be classified into four major groups: topological, geometric, electronic, and physicochemical. Topological descriptors were calculated using two-dimensional representation of the molecules. Geometric and electronic descriptors depend on the three-dimensional coordinates of atoms. Therefore, in order to calculate these types of descriptors one need to optimize the molecular structure of each compound. In the present work, the three-dimensional structure of each molecule was optimized using self-consistence molecular orbital method of AM1 (SCF-MO AM1) implemented in the MOPAC package (version 6) [15]. Some of the descriptors generated for each compound encoded similar information about the molecule of interest. Therefore, it was desirable to test each descriptor and eliminate those, which show high correlation ($R>$ 0.90) with each other. A total of 19 out of 77 descriptors showed high correlation and was removed from the consideration. Then, multiple linear regression (MLR) method was used to build the linear models that relate the RRFs to the structural parameters (descriptors). Three selected MLR models for FID, Ar-PID and Kr-PID are presented in Table 2.

3.3. STANN and ANN generation

The STANN and ANN programs were written in Fortran 77 in our laboratory. The networks were generated using the descriptors appearing in the MLR models as inputs. A three-layer network with a sigmoid transfer function was designed for each of STANNs and ANNs. Before training the STANNs and ANNs, the input and output values of the networks were normalized between 0.1 and 0.9. The number of nodes in the hidden layer, learning rate and momentum were optimized. The initial weights were selected randomly between -1 and $+1$. As can be seen from Table 2, the MLR model for FID RRFs includes seven descriptors. Therefore, the number of inputs in the STANN and ANN for FID were seven and the number of nodes in the output layer was set to be one. Two MLR models for Ar-PID and Kr-PID RRFs include six and five descriptors, respectively. Therefore, the number of inputs for the STANNs and ANNs for Ar-PID was six and in the case of Kr-PID was five, and the number of nodes in the output layer was set to be one. In order to evaluate the performance of the STANNs and ANNs, the standard error of training (SET) and the standard error of prediction (SEP) were used.

4. Results and discussion

As can be seen from Table 1, the data set consists of a diverse set of molecules. Table 2 shows the specifications of three selected MLR models for FID, Ar-PID, and Kr-PID RRFs. The mean effect for each parameter is also included in this table. Inspection of the variables appearing in the MLR models reveals that these parameters encode different aspects of the molecular structure and properties.

Table 3
Architectures of the STANNs and ANNs and specifications for different detection methods[a]

	FID		Ar-PID		Kr-PID	
	STANN	ANN	STANN	ANN	STANN	ANN
Number of nodes in the input layer	7	7	6	6	5	5
Number of nodes in the hidden layer	7	7	7	7	6	6
Number of nodes in the output layer	1	1	1	1	1	1
Number of iterations in the beginning of overfitting	295 000	472 500	18 000	18 000	83 000	319 000
Learning rate	0.8	0.8	0.99	0.99	0.9	0.9
Momentum	0.4	0.4	0.9	0.9	0.4	0.4

[a] The transfer function for all models is a sigmoid function.

In FID, the amount of ions formed determines the conductivity, which is registered as a response. The response of hydrocarbons in this detector is attributed to the number of carbon atoms from which they are made up and to the chemical nature of the molecules. The relative number of carbon atoms in the molecule (RENC) is an important descriptor that was appeared in the MLR model of FID. The FID response of heteroatom-substituted hydrocarbons is always less than that of the parent hydrocarbon. Therefore, the descriptors that contain information about the relative number or relative weight of carbon atoms in a compound are very important. RENC in the MLR model has a positive regression coefficient, and in agreement with the experiment indicates that as the number of carbon atoms increases the FID response increases. In addition, the process of response of organic structures in the FID starts with the thermal decomposition of C–X bonds, where X can be any atom. This can be represented by the number of different C–X bonds or by descriptors expressing the strength of such bonds. The presence of the descriptor of maximum bond order of carbon atoms (MBOC) in the MLR model can be attributed to this effect. The presence of descriptors expressing the strength of the X–H bonds such as maximum valency of hydrogen atoms (MVH) in the MLR model of FID are also very important. Since these bonds are on the branches of the molecules and are mostly exposed in molecular collisions, therefore the thermal cracking usually starts with these bonds. Appearance of the path one connectivity index ($^1\chi_p$) with negative mean effect in the MLR model of FID reveals that as degree of branching increases the RRF value decreases. Boiling point (b.p.) of the molecules as a physicochemical parameter with a relatively small negative mean effect of -0.338 is also appeared in the model. This is in agreement with the experiment, because as boiling point of a molecule is higher, the intermolecular forces are stronger and therefore, the ionization of these molecules in FID is more difficult. In order to improve the statistical parameters of the MLR models, different types of combination of descriptors such as square and cubic terms were examined. It can be seen from Table 2 that square of the polarizability (α^2) was entered in the MLR model for FID. However, addition of this parameter improves the results of the MLR model.

The second and third MLR models in Table 2 are due to Ar-PID and Kr-PID and have six and five descriptors, respectively. Inspection of these models reveals that topological, geometric and electronic properties of molecules play some roles in the mechanism of the relative response factors of these detectors. The PID response is directly proportional to the probability of photoionization and thus to the number of potentially photoionizable electrons in the molecule. The most important descriptor in the PID MLR models is energy of highest occupied molecular orbital (HOMO) which based on Koopmann theorem is numerically equal to the ionization potential (IP) but with a negative sign. Mean effect of this parameter in the models is negative indicating that a compound with a higher ionization potential (IP) has a lower value for the PID response. The presence of topological descriptors such as cluster three connectivity index ($^3\chi_c$) in the Ar-PID MLR model and path four connectivity index ($^4\chi_p$) in the Kr-PID MLR model with negative regression coefficients indicate that increasing of degree of branching decreases the PID RRFs. The other descriptor that is appeared in both MLR models is molecular density (MD). This parameter represents the ratio of molecular mass to the Van der Waals volume of the molecules that can be considered as a measure of the compactness of the molecules and therefore is a very important parameter affecting the PID response. It can be seen that mean effect of this parameter is negative for both Ar-PID and Kr-PID models. It is noteworthy that only the parameters of HOMO and MD are appeared in both MLR models for Ar-PID and Kr-PID systems while the remaining parameters are different. This is in agreement with the experiment, which shows that there is no correlation between the above-mentioned detectors [14].

The main goal of the present study was generation of the STANNs for modeling of FID and PID RRFs. In the case of STANN, a new method was used for updating the node's weights. Before the training of these networks, the parameters of the number of nodes in the hidden layer, learning rate and momentum were optimized. The procedure for the optimization of these parameters is reported in Refs. [4,5,16]. The architectures and specifications of the optimized STANNs for FID and PID systems are shown in Table 3. Also, a simple ANN was used for prediction of FID and PID RRFs. For comparison

Table 4
Experimental and calculated values of the RRFs for the training and prediction sets of FID, Ar-PID and Kr-PID using neural networks

No.[a]	FID			Ar-PID			Kr-PID		
	Exp	STANN	ANN	Exp	STANN	ANN	Exp	STANN	ANN
Training set									
1	0.1087	0.0706	0.0585	1.7177	1.7213	1.7114	0.1116	0.1133	0.1025
2	0.1660	0.1904	0.1964	0.6351	0.7349	0.7248	0.0118	0.0027	0.0123
3	0.2350	0.1974	0.1972	0.5497	0.5834	0.5868	0.0084	0.0889	0.0153
4	0.2577	0.2124	0.1962	1.0300	1.0796	1.0920	0.1352	0.1538	0.1359
5	0.2815	0.2666	0.2630	0.0273	0.0684	0.0637	0.0232	0.0632	0.0330
6	0.3106	0.3030	0.2944	0.9763	1.0497	1.0182	0.5900	0.6972	0.5690
7	0.3232	0.3416	0.3371	0.7868	0.8020	0.7883	0.4363	0.4923	0.3180
8	0.3234	0.3166	0.3386	2.0094	2.0519	2.0534	2.1856	2.1898	2.1773
9	0.3259	0.3518	0.3470	0.7507	0.8035	0.7828	0.2336	0.3070	0.2510
10	0.3334	0.3555	0.3436	0.7946	0.8279	0.8282	0.3357	0.3036	0.4183
11	0.3466	0.3245	0.3229	1.2989	1.3681	1.3597	1.1322	1.2420	1.1218
12	0.3587	0.4234	0.4294	0.5674	0.5584	0.5429	0.0406	−0.0020	−0.0311
13	0.3600	0.2794	0.2958	1.0752	1.1359	1.1196	0.0129	−0.0297	−0.0742
14	0.3647	0.3857	0.3856	0.7975	0.7787	0.7617	0.1299	0.1157	0.0496
15	0.3690	0.3315	0.3234	0.7558	0.7530	0.7333	0.1134	0.1053	0.0780
16	0.3787	0.3987	0.3964	0.5381	0.4542	0.4430	0.2584	0.3466	0.3098
17	0.3788	0.3787	0.3641	0.6644	0.6004	0.5990	0.0427	0.0274	0.0161
18	0.4073	0.4327	0.4282	0.6374	0.6595	0.6640	0.0832	0.0619	0.0680
19	0.4141	0.3452	0.3417	0.7357	0.8404	0.8137	0.0306	0.1191	0.0470
20	0.4158	0.4268	0.4280	0.6796	0.6867	0.6214	0.0344	0.0646	0.0870
21	0.4261	0.4130	0.4107	0.5213	0.4540	0.4448	0.1976	0.2713	0.2105
22	0.4270	0.4482	0.4434	0.5786	0.5928	0.6072	0.1315	0.1254	0.1592
23	0.4297	0.3801	0.3997	1.0928	1.0926	1.0742	0.6477	0.6161	0.6820
24	0.4344	0.3616	0.3576	0.7885	0.7668	0.7405	0.4555	0.5093	0.5212
25	0.4380	0.4187	0.4127	0.9715	0.8836	0.8758	0.1398	0.1473	0.1512
26	0.4449	0.4166	0.4177	0.4644	0.5019	0.4783	0.0436	0.1009	0.0759
27	0.4470	0.4626	0.4639	1.0463	0.8718	0.8564	0.6376	0.6399	0.6625
28	0.4569	0.4355	0.4335	0.7367	0.7057	0.7078	0.0126	0.0282	−0.0103
29	0.4575	0.4105	0.4030	0.7599	0.7410	0.7433	0.2215	0.2386	0.2489
30	0.4637	0.4540	0.4578	0.6098	0.5459	0.5334	0.1625	0.1340	0.1576
31	0.4720	0.4101	0.4117	0.5068	0.4662	0.4435	0.0650	0.1368	0.0680
32	0.4743	0.5261	0.5232	0.6321	0.7579	0.7267	0.6379	0.3394	0.6305
33	0.4748	0.4601	0.4440	0.8169	0.8263	0.7974	0.0514	0.0333	0.0635
34	0.4751	0.4852	0.4941	0.5420	0.4847	0.4684	0.0447	0.0792	0.0376
35	0.4814	0.4970	0.4935	0.7240	0.7723	0.7236	0.0375	−0.0583	0.0092
36	0.4827	0.4545	0.4499	1.0508	1.1537	1.1726	0.8419	0.7967	0.8545
37	0.4832	0.4428	0.4372	0.7992	0.8491	0.8278	ND	0.7353	0.7037
38	0.4868	0.4819	0.4823	0.8381	0.8029	0.7890	0.3257	0.3397	0.3271
39	0.4870	0.5017	0.5036	0.8727	0.7579	0.7433	0.3200	0.2930	0.2923
40	0.4941	0.4833	0.4745	0.7240	0.5887	0.5851	0.0075	0.0261	−0.0125
41	0.4959	0.5554	0.5582	0.5095	0.5369	0.5188	0.0630	0.0311	0.0531
42	0.4966	0.5014	0.5023	0.7385	0.7534	0.7382	0.3162	0.4756	0.4468
43	0.5007	0.4934	0.4926	0.3498	0.4761	0.4618	0.3908	0.2297	0.2088
44	0.5028	0.4565	0.4585	0.8149	0.8003	0.7835	0.1599	0.1531	0.2027
45	0.5041	0.4872	0.4843	1.5510	1.4763	1.4702	1.0475	1.0152	1.0165
46	0.5082	0.4564	0.4582	0.8102	0.5188	0.5129	0.1590	0.3256	0.3301
47	0.5166	0.5040	0.5028	0.5395	0.4833	0.4669	0.1443	0.2386	0.2630
48	0.5175	0.5211	0.5244	0.5996	0.5125	0.4895	0.0637	0.0130	0.0239
49	0.5200	0.4540	0.4527	0.4277	0.4499	0.4337	0.1962	0.2702	0.2876

Table 4. Continued

No.[a]	FID			Ar-PID			Kr-PID		
	Exp	STANN	ANN	Exp	STANN	ANN	Exp	STANN	ANN
50	0.5287	0.5117	0.5077	0.5879	0.5542	0.5051	0.1013	0.1381	0.1400
51	0.5316	0.4788	0.4814	0.6284	0.7285	0.7189	0.0727	0.0822	0.1315
52	0.5396	0.5306	0.5270	0.7517	0.7497	0.7325	0.3250	0.3073	0.3015
53	0.5449	0.5529	0.5667	0.4712	0.4976	0.4811	0.0440	0.1546	0.1219
54	0.5450	0.4733	0.4745	0.4573	0.4605	0.4461	0.2632	0.2755	0.2305
55	0.5462	0.5515	0.5485	0.6857	0.7757	0.7609	0.2971	0.3513	0.3338
56	0.5480	0.5586	0.5446	0.5724	0.5549	0.5455	0.1539	0.2172	0.1477
57	0.5563	0.4978	0.4898	0.8955	0.9174	0.8960	ND	0.6869	0.6702
58	0.5702	0.4869	0.4900	0.9096	0.8922	0.8805	0.2744	0.3086	0.2557
59	0.5721	0.4290	0.4274	0.4717	0.4561	0.4352	0.1393	0.1692	0.1159
60	0.5873	0.5770	0.5704	1.0451	0.9602	0.9499	0.5755	0.6235	0.6056
61	0.5898	0.6339	0.6159	0.7624	0.8204	0.7900	0.1730	0.1578	0.2062
62	0.6013	0.5243	0.5299	0.5767	0.5457	0.5657	0.0826	0.0296	0.0136
63	0.6013	0.5691	0.5670	0.7668	0.7218	0.7081	0.2845	0.3312	0.3162
64	0.6243	0.7080	0.6957	0.9569	0.7688	0.7465	0.7304	0.5791	0.5915
65	0.6458	0.5867	0.5959	0.6149	0.6004	0.6286	0.0660	0.0764	0.0535
66	0.6479	0.6014	0.6133	0.5562	0.4966	0.4843	0.2534	0.2518	0.2225
67	0.6483	0.6690	0.6603	1.0434	1.0055	0.9853	0.7115	0.6843	0.6810
68	0.6543	0.5039	0.5027	0.2055	0.1888	0.1744	0.0305	0.0200	−0.0289
69	0.6595	0.6491	0.6442	0.8801	0.7183	0.6972	0.5860	0.4612	0.4571
70	0.6646	0.6319	0.6279	0.8325	0.7118	0.7004	0.5486	0.5474	0.5528
71	0.6708	0.6687	0.6607	0.6971	0.6559	0.6349	0.3784	0.3582	0.3567
72	0.6762	0.6777	0.6717	0.6362	0.6903	0.6678	0.2983	0.3352	0.3267
73	0.6769	0.6113	0.6219	0.9675	0.9118	0.8926	1.2091	1.1232	1.0940
74	0.6809	0.6207	0.6220	0.7826	0.6964	0.6813	0.2856	0.3286	0.3419
75	0.6887	0.6302	0.6342	1.0316	0.7920	0.7662	0.3668	0.3289	0.3245
76	0.7009	0.7906	0.7762	0.5102	0.4985	0.4811	0.2665	0.1980	0.2504
77	0.7014	0.6825	0.6749	0.7449	0.6405	0.6149	0.3314	0.3149	0.3123
78	0.7115	0.7359	0.7225	0.5626	0.4677	0.4374	0.2879	0.1922	0.1878
79	0.7119	0.6857	0.6912	0.4564	0.5017	0.4841	0.0453	0.0528	0.0305
80	0.7158	0.6783	0.6696	0.0163	−0.0095	0.0011	0.0111	−0.0013	0.0139
81	0.7242	0.6879	0.6784	0.7484	0.6538	0.6398	0.0184	0.0094	0.0961
82	0.7272	0.7405	0.7309	0.5662	0.5157	0.4992	0.4715	0.3280	0.2480
83	0.7298	0.6942	0.6691	0.3751	0.4715	0.4530	0.4287	0.2402	0.2828
84	0.7343	0.6987	0.7015	0.4853	0.5798	0.5634	0.0406	−0.0220	0.0308
85	0.7350	0.6863	0.6885	0.7993	0.9537	0.9119	0.5923	0.5986	0.6568
86	0.7589	0.7350	0.7176	0.8594	0.7483	0.7382	0.0719	0.0541	0.0304
87	0.7678	0.7293	0.7493	0.9855	0.8359	0.8253	0.4068	0.3439	0.3986
88	0.7689	0.7419	0.7386	0.7031	0.6764	0.6537	0.0450	0.0296	0.0705
89	0.8058	0.7734	0.7619	0.6309	0.5541	0.5505	0.0254	0.0295	0.0502
90	0.8423	0.8343	0.8311	0.5681	0.6688	0.6394	0.4893	0.4892	0.4726
91	0.8642	0.8756	0.8718	0.9582	0.8978	0.8922	0.3190	0.2929	0.2751
92	0.8736	0.8752	0.8828	0.7495	0.6566	0.6402	0.3627	0.2941	0.2964
93	0.9204	0.8529	0.8591	0.8025	0.7352	0.7217	0.3477	0.3859	0.3665
94	0.9358	0.9539	0.9461	0.5808	0.5291	0.5328	0.0621	−0.0027	−0.0074
95	0.9371	0.6595	0.6523	0.8858	0.7674	0.7552	0.6240	0.4535	0.4484
96	0.9440	0.7376	0.7416	0.4111	0.4541	0.4272	0.0804	0.0302	0.0059
97	0.9473	0.9680	0.9745	0.0354	−0.0163	−0.0331	0.0450	0.0328	0.0054
98	0.9527	0.9808	0.9557	0.6832	0.7158	0.6975	0.0773	0.1800	0.2528
99	1.0062	0.9011	0.9283	0.8109	0.8493	0.838	0.0278	0.0351	0.0247
100	1.0232	1.0286	1.0199	0.7023	0.6238	0.6051	0.0404	0.0013	0.0732
101	1.1259	1.1264	1.1225	0.6744	0.7418	0.7268	0.4911	0.5012	0.5139

Table 4. Continued

No.[a]	FID			Ar-PID			Kr-PID		
	Exp	STANN	ANN	Exp	STANN	ANN	Exp	STANN	ANN
102	1.1690	1.1077	1.0994	1.2599	1.1965	1.1798	0.7967	0.7051	0.7323
103	1.2200	1.2205	1.2158	0.8260	0.7759	0.7601	0.0334	0.0488	0.0919
104	1.2267	1.1801	1.1764	0.7407	0.7464	0.7346	0.3986	0.4072	0.4206
105	1.2935	1.2821	1.2775	1.0668	1.0802	1.0854	0.2015	0.3600	0.1990
Prediction set									
106	0.2779	0.3512	0.2904	0.8297	0.8220	0.7860	0.0262	0.0060	−0.0497
107	0.3548	0.3538	0.3667	0.8873	0.7211	0.7238	0.0115	−0.1036	−0.1065
108	0.4244	0.4690	0.4680	0.9059	0.8800	0.8351	0.5500	0.6383	0.5269
109	0.4292	0.4459	0.4367	0.6554	0.5564	0.5332	0.2239	0.2135	0.2183
110	0.4786	0.4206	0.4116	0.5495	0.5760	0.5780	ND	0.2332	0.3268
111	0.5112	0.4814	0.4814	0.6921	0.6939	0.6784	0.0820	0.0478	0.0624
112	0.5203	0.5215	0.5303	0.0963	0.0612	0.0777	0.0342	−0.0646	−0.0319
113	0.5412	0.4979	0.4887	0.8146	0.8806	0.8737	ND	0.3434	0.2612
114	0.5809	0.4982	0.5019	0.5823	0.7454	0.7036	0.0298	−0.022	0.0376
115	0.6143	0.6225	0.6190	0.5086	0.5720	0.5536	0.0396	0.0731	−0.0135
116	0.6502	0.7412	0.7427	0.6271	0.5408	0.5318	0.2283	0.2916	0.3146
117	0.6668	0.5993	0.5952	0.1082	0.2085	0.2675	0.0192	0.0243	−0.0043
118	0.6803	0.6566	0.6627	1.1692	0.9016	0.9054	1.3453	1.2534	1.1833
119	0.7200	0.6641	0.6554	0.7280	0.6949	0.6733	ND	0.3658	0.3571
120	0.7408	0.8443	0.8224	0.5744	0.5637	0.5506	0.3231	0.1878	0.2240
121	0.8462	0.7905	0.7782	0.4466	0.4663	0.4450	0.4157	0.2205	0.2772
122	1.1528	1.0842	1.0689	0.7001	0.7977	0.7697	0.4803	0.6645	0.7022

ND: Not detected.

[a] The numbers refer to the numbers of the molecules given in Table 1.

purposes, the architectures and specifications of the optimized FID and PID ANNs are also given in Table 3. In order to control the overfitting of the networks during the training procedure, the SET and SEP values were recorded after each 500 iterations. In the case of STANN, for FID after 295 000 iterations the values of SEP started to increase and overtraining began and for Ar-PID and Kr-PID, overtraining began after 18 000 and 83 000 iterations, respectively. These numbers should be compared with 472 500, 18 000, and 319 000, respectively, for the conventional ANN. This means that in agreement with our previous work the training of the STANN is much faster compared with that of the simple ANN. The number of inputs in the STANNs is the same as the number of descriptors appearing in the MLR models. The topology of STANN for FID was 7-7-1, and the structures of STANN for Ar-PID and Kr-PID were 6-7-1 and 5-6-1, respectively. For the evaluation of the prediction ability of the STANNs and ANNs, the trained STANNs and ANNs were used to predict the RRFs of the molecules included in the prediction set. The calculated values of RRFs using the generated STANNs and ANNs for the training and the prediction sets of FID, Ar-PID and Kr-PID are presented in Table 4. For Kr-PID, the RRFs of five compounds that their experimental values are not available were calculated using the STANN and ANN and are given in Table 4. The statistical parameters, such as correlation coefficients (R) between the calculated and experimental values of RRFs and standard errors (SEs) for the training and prediction sets obtained using the STANNs, ANNs, and MLR models are shown in Table 5. Inspections of the SET and SEP values for the STANN, ANN and MLR methods reveal the superiority of the neural networks over that of the MLR in predicting of the RRFs. This is due to the nonlinear capabilities of the STANNs and ANNs. As can be seen from Table 5, the values of the SET and

Table 5
Statistical parameters obtained using the STANN, ANN and MLR models for RRFs of different detection methods

Detection method	Model	SET (%)	SEP (%)	$R_{training}$	$R_{prediction}$
FID	STANN	4.107	3.849	0.975	0.959
	ANN	3.961	3.778	0.977	0.963
	MLR	12.051	8.240	0.859	0.889
Ar-PID	STANN	3.703	4.076	0.963	0.926
	ANN	3.735	4.161	0.964	0.930
	MLR	18.220	20.897	0.793	0.653
Kr-PID	STANN	2.680	3.798	0.974	0.959
	ANN	2.516	3.677	0.979	0.962
	MLR	19.937	29.836	0.794	0.610

SEP are comparable for each method and since this is true for all models one may conclude that the prediction set is a very good representative of the training set. However, it is common to have an additional external validation set to make sure that the predictive ability of the neural networks is good. Unfortunately, we were not been able to find a set of molecules with RRF values obtained exactly at the same condition and with RRF values of full range. Therefore, we have randomly chosen two different test sets each consisting of 17 molecules and the networks were trained using the remaining molecules. The results of these sets for the three detectors are given in Table 6. As can be seen from this table, the values of SET and SEP for the prediction set are similar to those of the test sets in the cases of STANN and ANN models. This indicates that the networks are independent from the prediction set. On the other hand, the values of SET and SEP that are between 2.497% to 4.861% indicating the good predictive ability of the neural networks in predicting of the RRFs.

Fig. 2a–c show the plots of the STANN calculated values of RRFs against the experimental values for FID, Ar-PID and Kr-PID, respectively. Fig. 3a–c show the plots of the residuals against the experimental values of RRFs, for the STANNs and ANNs models of FID, Ar-PID and Kr-PID. The

Table 6
Comparison of the SET and SEP of the two test sets with the prediction set for different detection methods

Detection method	Model	STANN		ANN	
		SET (%)	SEP (%)	SET (%)	SEP (%)
FID	Prediction set	4.107	3.849	3.961	3.778
	Test set I	4.008	3.721	4.048	3.854
	Test set II	4.671	3.913	4.699	3.970
Ar-PID	Prediction set	3.703	4.076	3.735	4.161
	Test set I	3.322	4.832	3.354	4.861
	Test set II	3.873	4.547	3.645	4.291
Kr-PID	Prediction set	2.680	3.798	2.516	3.677
	Test set I	2.813	4.383	2.497	4.673
	Test set II	2.574	3.961	2.696	3.927

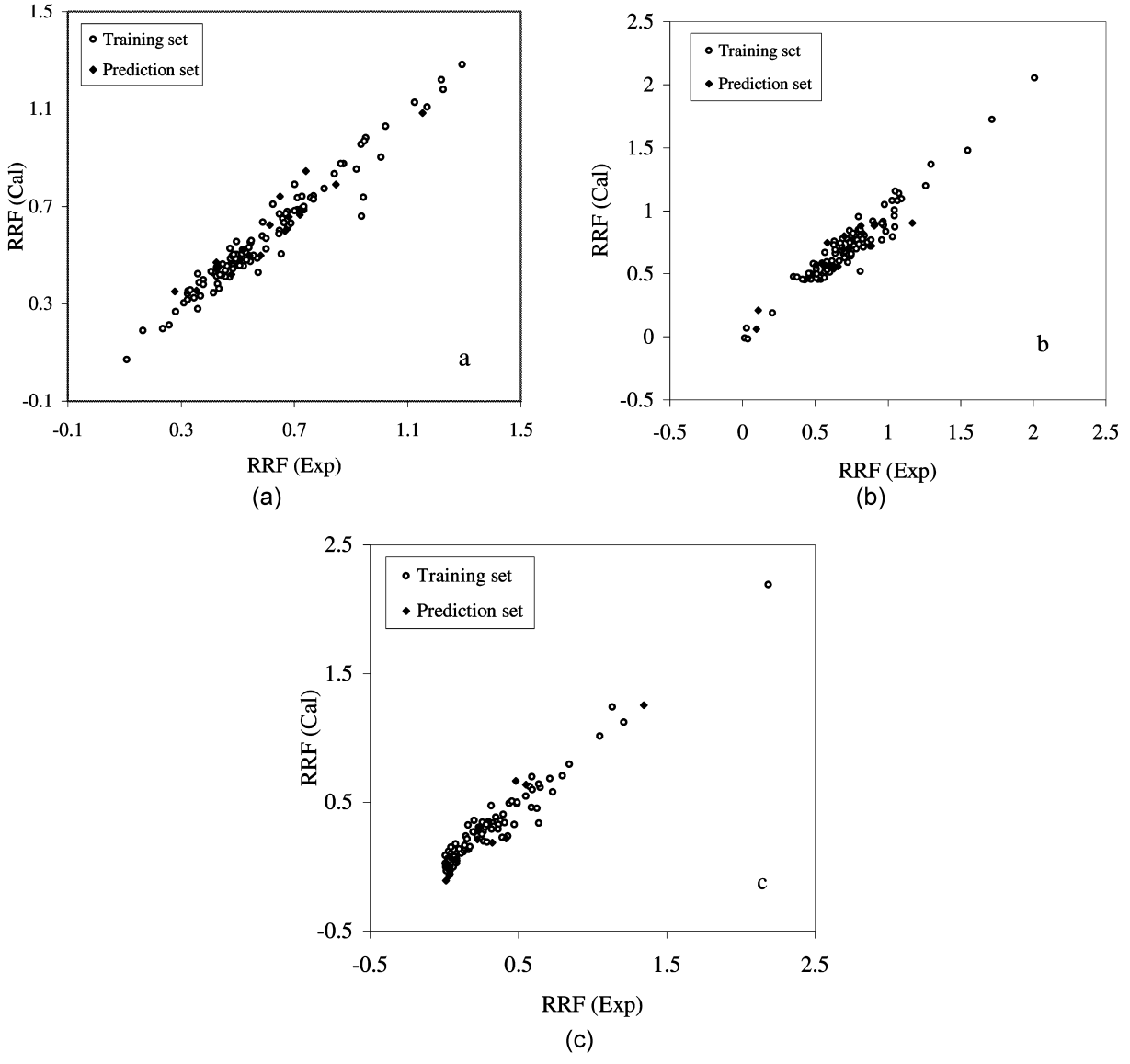

Fig. 2. Plot of the STANN calculated values of RRFs versus the experimental values (a) FID; (b) Ar-PID; (c) Kr-PID.

propagation of the residuals in both sides of zero indicate that no systematic error exist in the development of the STANNs and ANNs.

5. Conclusions

Comparison of the values of the SET and SEP obtained using different models of STANN, ANN and MLR for predicting of RRFs of different detectors shows superiority of the neural networks over that of linear regression model. Inspection of the statistical parameters (Table 5) indicates that the use of the linear models in predicting of RRFs of different detectors is not justified. Since the improvement of the results obtained using the nonlinear

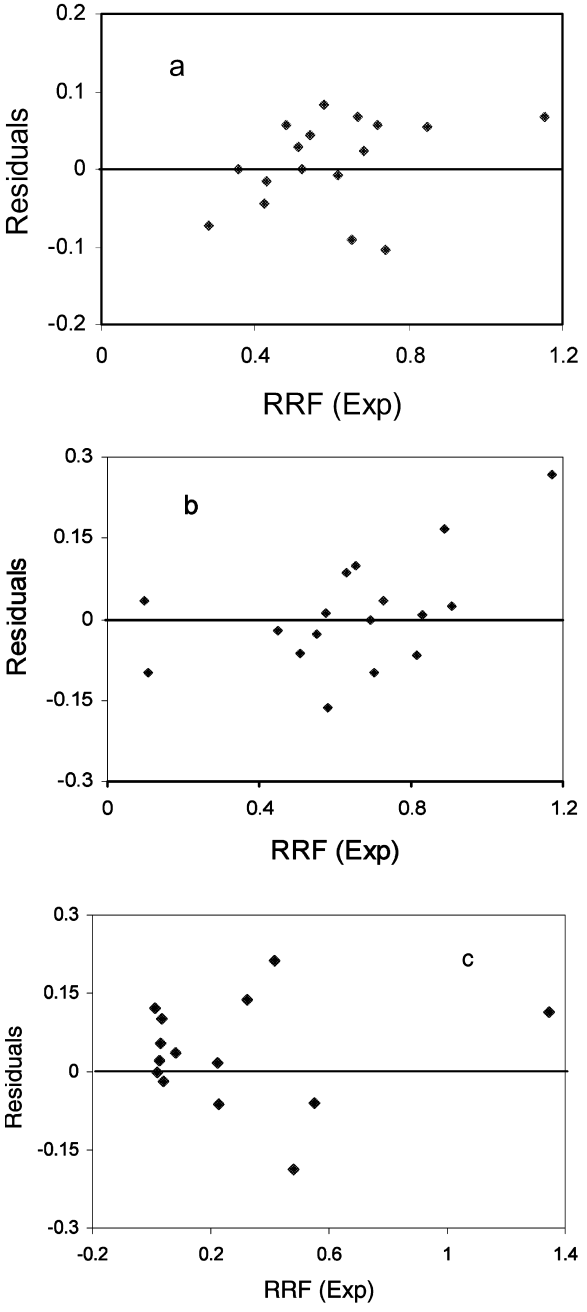

Fig. 3. Plot of STANN residuals versus the experimental values of RRFs for the prediction set (a) FID; (b) Ar-PID; (c) Kr-PID.

models of artificial neural networks is considerable, one may conclude that the nonlinear characteristics of the RRFs are serious.

Investigation of the calculated STANN values of RRFs with those of simple ANN indicates that these models are comparable in predicting FID, Ar-PID and Kr-PID RRFs for a variety of organic compounds. The only advantage of the STANN over that of the simple ANN is that the optimization procedure of the former is much faster. In addition, in contradiction to our previous conclusion [13], the numbers of adjustable parameters for both models are the same and therefore, there is no difference between the validity of these models in prediction of the RRFs.

References

[1] A.R. Katritzky, E.V. Gordeeva, J. Chem. Inf. Comput. Sci. 33 (1993) 835.
[2] A.R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, M. Karelson, Anal. Chem. 66 (1994) 1799.
[3] B. Lucic, N. Trinajstic, S. Sild, M. Karelson, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 39 (1999) 610.
[4] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 825 (1998) 161.
[5] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. A 897 (2000) 227.
[6] W.E. Wentworth, Y. Li, S.D. Stearns, J. High Resolut. Chromatogr. 19 (1996) 85.
[7] G. Gremaud, W.E. Wentworth, A. Zlatkis, R. Swatloski, E.C.M. Chen, S.D. Stearns, J. Chromatogr. A 724 (1996) 235.
[8] W.E. Wentworth, S. Wiedemann, Y. Qin, J. Madabushi, S.D. Steams, J. Appl. Spectrosc. 49 (9) (1995) 1282.
[9] D.W. Patterson, Artificial Neural Networks: Theory and Applications, Simon and Schuster, New York, 1996.
[10] J. Zupan, J. Gasteiger, Anal. Chim. Acta 248 (1991) 1.
[11] N.K. Bose, P. Liang, Neural Network—Fundamentals, McGraw-Hill, New York, 1996.
[12] http://www.imagination-engines.com/adpcstanno.htm
[13] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A, in press.
[14] W.E. Wentworth, N. Helias, A. Zlatkis, E.C.M. Chen, S.D. Stearns, J. Chromatogr. A 795 (1998) 319.
[15] MOPAC Package, Version 6, US Air Force Academy, Colorado Springs, CO, 80840.
[16] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. A 927 (2001) 211.