Prediction Partial Molar Heat Capacity at Infinite Dilution for Aqueous Solutions of Various Polar Aromatic Compounds over a Wide Range of Conditions Using Artificial Neural Networks

Aziz Habibi-Yangjeh* **and Mahdi Esmailian**

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardabili, P.O. Box 179, Ardabil, Iran
**E-mail: ahabibi@uma.ac.ir; habibiyangjeh@yahoo.com*

Received December 21, 2006

Artificial neural networks (ANNs), for a first time, were successfully developed for the prediction partial molar heat capacity of aqueous solutions at infinite dilution for various polar aromatic compounds over wide range of temperatures (303.55-623.20 K) and pressures (0.1-30.2 MPa). Two three-layered feed forward ANNs with back-propagation of error were generated using three (the heat capacity in T = 303.55 K and P = 0.1 MPa, temperature and pressure) and six parameters (four theoretical descriptors, temperature and pressure) as inputs and its output is partial molar heat capacity at infinite dilution. It was found that properly selected and trained neural networks could fairly represent dependence of the heat capacity on the molecular descriptors, temperature and pressure. Mean percentage deviations (MPD) for prediction set by the models are 4.755 and 4.642, respectively.

Key Words : Artificial neural networks, Partial molar heat capacity, Aqueous solutions, Polar aromatic compounds, Theoretical descriptors

Introduction

Heat capacities of organic solutes in water are of great interest for calculating thermodynamic properties of organic aqueous systems at super ambient conditions. The temperature integration of the heat capacity data allows obtaining the standard chemical potentials and activity coefficients needed for calculating phase and chemical equilibria at conditions of interest for geochemistry, power cycle chemistry and hydrothermal technologies.[1,2] Only a limited amount of data are available at upper temperatures and pressures. The main reason is certainly a time consuming and costly task of the calorimeter construction since commercial instruments allowing the heat capacity determination over a range of temperatures do not have the precision necessary for the calculation of heat capacity. For this reason, it is very valuable to predict the heat capacity at higher temperatures and pressures using minimum number of experiments. The prediction of physicochemical and biological properties/activities for organic molecules is the main objectives of the quantitative structure-property/activity relationships (QSPRs/QSARs).[3-9] QSPR/QSAR models are obtained on the basis of the correlation between the experimental values of the property/activity and descriptors reflecting the molecular structure of the respective compounds. Since these theoretical descriptors are determined solely from computational methods, a priori predictions of the properties/activities of compounds are possible, no laboratory measurements are needed thus saving time, space, materials, equipment and alleviating safety (toxicity) and disposal concerns.[10,11]

Various methods for constructing QSPR/QSAR models have been used including multi-parameter linear regression (MLR), principal component analysis (PCA) and partial least-squares regression (PLS).[12-15] In addition, artificial neural networks (ANNs) have become popular due to their success where complex non-linear relationships exist amongst data.[16-18] ANNs are biologically inspired computer programs designed to simulate the way in which the human brain processes information.[18] ANNs gather their knowledge by detecting the patterns and relationships in data, not from programming. The wide applicability of ANNs stems from their flexibility and ability to model non-linear systems without prior knowledge of an empirical model. For these reason in recent years, ANNs have been used to a wide variety of chemical problems such as simulation of mass spectra, ion interaction chromatography, aqueous solubility and partition coefficient, simulation of nuclear magnetic resonance spectra, prediction of bioconcentration factor, solvent effects on reaction rate, prediction normalized polarity parameter in mixed solvent systems, acidity constant of organic compounds and dielectric constant of binary mixtures.[19-41]

In this work an ANN model, for a first time, was generated for prediction partial molar heat capacity of aqueous solutions at infinite dilution for various polar aromatic compounds over wide range of temperatures (303.55-623.20 K) and pressures (0.1-30.2 MPa) using three inputs (the partial molar heat capacity at infinite dilution for the various aqueous solutions at T = 303.55 K and P = 0.1 MPa, temperature and pressure). In the next step, a MLR model was constructed between the heat capacity of the compounds and four theoretical descriptors. Then an ANN model using the theoretical descriptors, temperature and pressure was constructed for prediction the heat capacity and the results were compared with the experimental values of them.

Methods and Procedure

Data set. A reliable database is critically important for the training of ANNs. Very recently partial molar heat capacity at infinite dilution have been determined for different aqueous solutions of polar aromatic compounds at various temperatures and pressures.[1,2] In this work, the data for aqueous solutions of phenol, *o*-cresol, *m*-cresol, *p*-cresol, aniline, *o*-toluidne, *m*-toluidine, *p*-toluidine, m-aminophenol and *o*-diaminobenzene that they have at least eight values for the heat capacity at various temperatures and pressures have been used as data set. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 74, 21 and 21 data, respectively.[18,19] The training and validation sets were used for the model generation and the prediction set was used for evaluation of the generated model, because a prediction set is a better estimator of the ANN generalization ability than a monitoring (validation) set.

Descriptor generation. In order to calculate the theoretical descriptors, the z-matrices (molecular models) were constructed with the aid of HyperChem 7.0 and molecular structures were optimized using AM1 algorithm.[42] In order to calculate theoretical descriptors, the molecular geometries of molecules were further optimized by *Dragon* package version 2.1.[43] For this purpose the output of the HyperChem software for each compound fed into the *Dragon* program and the descriptors were calculated. As a result, a total of 1481 theoretical descriptors were calculated for each compound in the data sets (11 compounds).

Feature selection. The theoretical descriptors were reduced by the following methods: 1) descriptors that are constant or nearly constant have been eliminated, because these descriptors can not define the variation of the property with structure; 2) in order to decrease the redundancy existing in the descriptors data matrix, the correlation coefficients for all pairs of remaining descriptors were determined. If a correlation coefficient was higher than 0.91, the descriptor with lower correlation with the heat capacity was eliminated;[44,45] 3) the method of stepwise multi-parameter linear regression was used to select the most important descriptors and to calculate the coefficients relating the heat capacity to the descriptors.[15] The MLR models were generated using spss/pc software package release 10.0.[46]

Neural network generation. The specification of a typical neural network model requires the choice of the type of inputs, the number of hidden layers, the number of neurons in each hidden layer and the connection structure between the inputs and the output layers. Three-layer networks with sigmoidal transfer function for neurons were designed. The initial weights were randomly selected between 0 and 1. Before training, the input and output values were normalized between 0.1 and 0.9. The optimization of the weights and biases was carried out according to the resilient back-propagation algorithm.[46] For evaluation predictive power of the networks, the trained ANNs were used to predict the heat capacity for 21 aqueous solutions included in the prediction set. The performances of ANNs are evaluated by the mean percentage deviation (MPD) and root-mean square error (RMSE), which are defined as follows:

$$\mathrm{MPD} = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{(P_i^{calc} - P_i^{\exp})}{P_i^{\exp}}\right| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(P_i^{calc} - P_i^{\exp})^2}{N}} \tag{2}$$

where $P_i^{\exp}$ and P_i^{cal} are experimental and calculated values of the heat capacity using the models.

Individual percent deviation (IPD) is defined as follows:

$$\mathrm{IPD} = 100 \times \left(\frac{P_i^{calc} - P_i^{\exp}}{P_i^{\exp}}\right) \tag{3}$$

The processing of the data was carried using Matlab 6.5.[47] The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab.[48]

Results and Discussion

Prediction the heat capacity without theoretical descriptors. There are no theoretical principles for choosing the proper network topology; so different structures were tested in order to obtain the optimal hidden neurons and training cycles.[18,19] Before training the network, the numbers of nodes in the hidden layer were optimized. In order to optimize the number of nodes in the hidden layer, several training sessions were conducted with different numbers of hidden nodes (from one to eleven). The root mean squared error of training (RMSET) and validation (RMSEV) sets were plotted *versus* the number of iterations for different number of neurons at the hidden layer and the minimum value of RMSEV was recorded as the optimum value. Plot of RMSET and RMSEV *versus* the number of nodes in the hidden layer has been demonstrated in Figure 1. It is clear that nine nodes in hidden layer is optimum value.

This network consists of three inputs including the partial molar heat capacity at infinite dilution for the various aqueous solutions (at T = 303.55 K and P = 0.1 MPa), temperature and pressure. Then an ANN with architecture 3-9-1 was generated. It is note worthy that training of the network was stopped when the RMSEV started to increases i.e. when overtraining begins. The overtraining causes the ANN to loose its prediction power.[32] Therefore, during training of the networks, it is desirable that iterations are stopped when overtraining begins. To control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations.

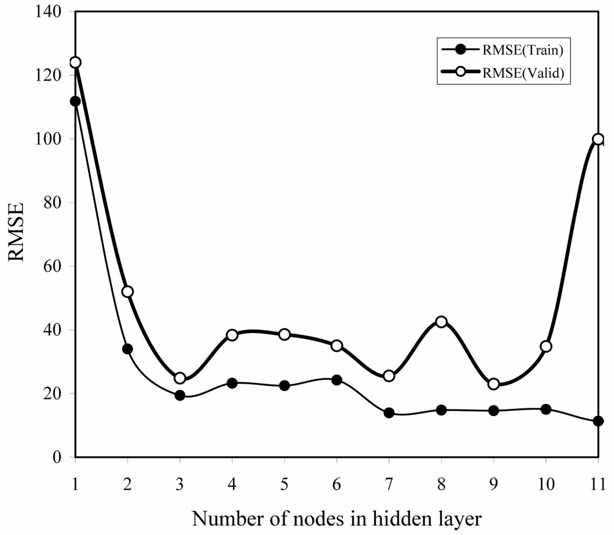

Figure 1. Plot of RMSE for training and validation sets *versus* the number of nodes in hidden layer.

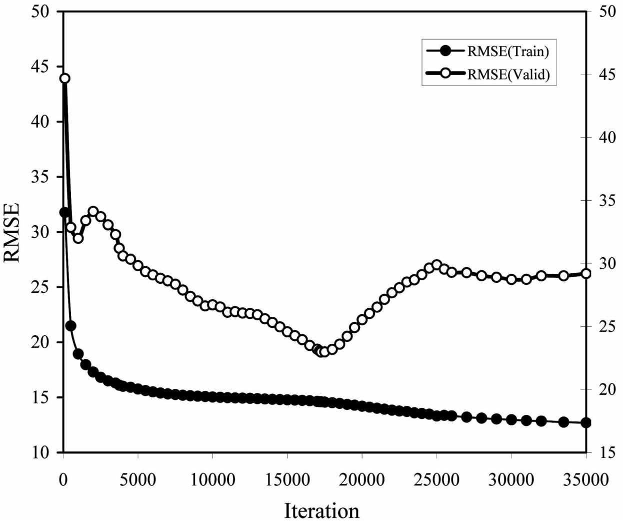

Figure 2. Plot of RMSE for training and validation sets (for the ANN model with architecture 3-9-1) *versus* the number of iterations.

Results obtained showed that after 17250 iterations the value of RMSEV started to increase and overfitting began (Figure 2).

The generated ANN was then trained using the training set for optimization of the weights and biases. For evaluation predictive power of the generated ANN, an optimized network was applied for prediction the heat capacity of different aqueous solutions at various temperatures and pressures in the prediction set, which were not used in the modeling procedure. Values of partial molar heat capacity for different aqueous solutions of various polar aromatic compounds along with the calculated and IPD values at various temperatures and pressures for training, validation and prediction sets have been shown in Table 1.

The correlation equation for all of the calculated values of the heat capacity from the ANN model and the experimental values is as follows:

$$C^{o}_{p,2}\,(\text{cal}) = 0.9606\ C^{o}_{p,2}\,(\text{exp}) - 13.849 \qquad (4)$$

$$N - 116;\ R - 0.9859;\ MPD - 3.017;$$
$$RMSE - 19.642;\ F - 3950.35$$

Similarly, the correlation of $C^{o}_{p,2}$ (cal) values *versus* $C^{o}_{p,2}$ (exp) in prediction set gives equation (5):

$$C^{o}_{p,2}\,(\text{cal}) - 0.9899\ C^{o}_{p,2}\,(\text{exp}) - 11.531 \qquad (5)$$

$$N = 21;\ R = 0.9761;\ MPD = 4.755;$$
$$RMSE = 29.179;\ F = 383.29$$

As can be seen the calculated values of the heat capacity are in good agreement with those of the experimental values. Plot of IPD for $C^{o}_{p,2}$ values in prediction set *versus* the experimental values of it has been illustrated in Figure 3. The results demonstrate that the MPD value for $C^{o}_{p,2}$ values in the prediction set is 4.755. As can be seen the model did not show proportional and systematic error, because the propagation of errors in both sides of zero are random (Figure 3).

The correlation coefficient (R), RMSE, MPD and statistical F-value of the model for total, training, validation and prediction sets show potential of the ANN model for simulation the complicated nonlinear relationship between the partial molar heat capacity at infinite dilution for aqueous solutions of the various polar aromatic compounds on the heat capacity in T − 303.55 K and P − 0.1 MPa, temperature and pressure (Table 2).

Prediction the heat capacity using theoretical descriptors. After feature selection (see section methods and procedure), multi-parameter linear correlation of the heat capacity *versus* the molecular descriptors in the training set gives the results in Table 3. It can be seen that four descriptors are appeared in the MLR model. These descriptors are: complementary information content (neighborhood symmetry of 0-order) (CIC0), geary autocorrelation-lag3/weighted by atomic masses (GATS3m), radia distribution function-5.0/weighted by atomic masses (RDF050m) and 3D-MoRSF-signal 08/weighted by atomic polarizabilities (Mor08p).

The correlation equation for the calculated values of $C^{o}_{p,2}$ *versus* the experimental values is as follows:

$$C^{o}_{p,2}\,(\text{cal}) = 0.9993\ C^{o}_{p,2}\,(\text{exp}) - 0.2194 \qquad (6)$$

$$N - 10;\ R - 0.99968;\ MPD - 0.1819;$$
$$RMSE - 0.7938;\ F - 21880.83$$

The next step in this work is the generation of the ANN model using theoretical descriptors. The artificial neural network consists of six inputs (including four descriptors appearing in the MLR model, temperature and pressure) and one output for $C^{o}_{p,2}$. Plot of RMSET and RMSEV *versus* the number of nodes in the hidden layer has been demonstrated in Figure 4. It is clear that three nodes in hidden layer is optimum value.

Then an ANN with architecture 6-3-1 was generated. To

Table 1. Values of partial molar heat capacity at infinite dilution for aqueous solutions of various polar aromatic compounds along with the calculated and IPD (individual percent deviation) at various temperatures and pressures using the ANN models

No.	Aqueous solutions	Data set	T	P	$C^{o}_{p,2}$ (exp)	$C^{o}_{p,2}$ (cal)[a]	IPD[a]	$C^{o}_{p,2}$ (cal)[b]	IPD[b]
1	phenol	training	303.55	0.1	325.1	313.19	−3.66	318.06	−2.17
2	phenol	validation	372.23	2.2	312.3	320.64	2.67	325.15	4.11
3	phenol	training	422.61	2.1	314.8	323.88	2.88	334.24	6.18
4	phenol	training	473.54	2.1	327.6	329.08	0.45	347.02	5.93
5	phenol	prediction	523.52	5	364.7	359.64	−1.39	366.56	0.51
6	phenol	training	573.38	10	467.5	458.76	−1.87	454.17	−2.85
7	phenol	training	574.58	10.2	463.4	463.35	−0.01	458.31	−1.10
8	phenol	validation	598.22	13.3	626.8	581.99	−7.15	615.72	−1.77
9	phenol	training	524.37	10.1	354.4	360.9	1.83	354.09	−0.09
10	phenol	training	573.3	20.2	382.7	378.46	−1.11	375.34	−1.92
11	phenol	training	598.17	20.2	463.9	469.32	1.17	468.58	1.01
12	phenol	training	474.89	30.3	311	323.5	4.02	316.75	1.85
13	phenol	training	523.6	30.1	324.1	326.23	0.66	323.42	−0.21
14	phenol	validation	573.22	30	346.9	349.56	0.77	342.33	−1.32
15	phenol	training	573.95	30	350.6	350.17	−0.12	342.89	−2.20
16	phenol	training	598.2	30.4	374.9	371.1	−1.01	374.58	−0.09
17	phenol	prediction	623.2	30.2	446.8	397.94	−10.94	497.17	11.27
18	*o*-cresol	training	303.55	0.1	407	406.84	−0.04	400.39	−1.62
19	*o*-cresol	validation	372.23	2.2	384	402.77	4.89	405.78	5.67
20	*o*-cresol	training	422.61	2.1	389.6	397.15	1.94	409.38	5.08
21	*o*-cresol	training	473.53	2.1	416.5	406.96	−2.29	415.31	−0.29
22	*o*-cresol	prediction	523.51	5	461.9	441.38	−4.44	444.79	−3.70
23	*o*-cresol	training	573.39	10.1	623.3	624.79	0.24	642.81	3.13
24	*o*-cresol	training	574.28	10.1	616	630.64	2.38	652.91	5.99
25	*o*-cresol	training	598.22	13.3	939.3	889.62	−5.29	898.77	−4.32
26	*o*-cresol	training	524.37	10.1	439.9	451.93	2.73	432.55	−1.67
27	*o*-cresol	training	573.3	20.2	513.9	513.44	−0.09	497.34	−3.22
28	*o*-cresol	prediction	598.17	20.3	676.9	658	−2.79	702.49	3.78
29	*o*-cresol	training	474.89	30.3	383.6	386.4	0.73	398.96	4.00
30	*o*-cresol	training	523.6	30.3	397.1	398.5	0.35	407.29	2.57
31	*o*-cresol	validation	573.23	30.1	411.4	447.07	8.67	441.72	7.37
32	*o*-cresol	training	574.04	30	436.2	448.91	2.91	443.46	1.67
33	*o*-cresol	training	598.2	30.4	558.6	519.84	−6.94	520.12	−6.89
34	*o*-cresol	prediction	623.2	30.1	747.9	670.06	−10.41	786.70	5.19
35	*m*-cresol	training	303.55	0.1	397.3	394.61	−0.68	389.87	−1.87
36	*m*-cresol	validation	372.24	2.2	363.8	392.58	7.91	398.54	9.55
37	*m*-cresol	training	422.62	2.1	385.8	395.54	2.52	404.36	4.81
38	*m*-cresol	training	473.53	2.1	399.5	416.8	4.33	410.81	2.83
39	*m*-cresol	training	523.51	5	453.8	448.36	−1.2	432.17	−4.77
40	*m*-cresol	training	573.38	10	587.5	579.85	−1.3	568.24	−3.28
41	*m*-cresol	training	574.48	10.2	587.2	586.69	−0.09	573.94	−2.26
42	*m*-cresol	validation	598.22	13.3	830.2	817.25	−1.56	784.74	−5.48
43	*m*-cresol	training	524.38	10.1	431.9	420.37	−2.67	422.76	−2.12
44	*m*-cresol	training	573.3	20.2	463	451.64	−2.45	465.86	0.62
45	*m*-cresol	prediction	598.17	20.3	612.4	586.86	−4.17	608.29	−0.67
46	*m*-cresol	training	474.89	30.4	373.7	384.75	2.96	387.24	3.62
47	*m*-cresol	training	523.6	30.3	387.2	396.01	2.28	397.79	2.74
48	*m*-cresol	validation	573.23	30.1	430.5	440.45	2.31	425.09	−1.26
49	*m*-cresol	training	573.75	29.8	433.5	441.82	1.92	426.58	−1.60
50	*m*-cresol	prediction	598.2	30.2	490.7	508.79	3.69	480.45	−2.09
51	*m*-cresol	training	623.17	30.2	647.9	648.69	0.12	669.99	3.41
52	*p*-cresol	training	303.55	0.1	400.6	398.57	−0.51	395.41	−1.30
53	*p*-cresol	validation	372.24	2.2	370.4	395.42	6.75	402.43	8.65
54	*p*-cresol	training	422.63	2.1	376.8	394.9	4.8	407.05	8.03
55	*p*-cresol	training	473.53	2.1	397.6	412.21	3.67	412.81	3.83
56	*p*-cresol	prediction	523.51	5	450.7	444.8	−1.31	435.46	−3.38
57	*p*-cresol	training	573.38	10	579.9	596.26	2.82	583.58	0.63
58	*p*-cresol	training	574.1	10.1	576.3	601.48	4.37	588.13	2.05
59	*p*-cresol	training	598.22	13.3	809.2	844.82	4.4	811.10	0.24

Table 1. Continued

No.	Aqueous solutions	Data set	T	P	$C^o_{p,2}$ (exp)	$C^o_{p,2}$ (cal)[a]	IPD[a]	$C^o_{p,2}$ (cal)[b]	IPD[b]
60	*p*-cresol	training	524.38	10.1	432.1	432.14	0.01	425.78	−1.46
61	*p*-cresol	training	573.3	20.2	459.8	469.55	2.12	473.14	2.90
62	*p*-cresol	prediction	598.18	20.4	595.6	604.82	1.55	625.71	5.05
63	*p*-cresol	training	474.89	30.4	367.2	385.33	4.94	393.29	7.11
64	*p*-cresol	training	523.6	30.3	383.6	396.88	3.46	402.51	4.93
65	*p*-cresol	validation	573.23	30.3	423	442.46	4.6	429.82	1.61
66	*p*-cresol	training	573.78	29.9	420.7	444.19	5.58	431.68	2.61
67	*p*-cresol	training	598.2	30.4	488	512.12	4.94	488.55	0.11
68	*p*-cresol	prediction	623.17	29.9	622.8	658.76	5.77	702.39	12.78
69	aniline	training	303.55	0.1	336.6	342.3	1.69	321.44	−4.51
70	aniline	validation	372.24	2.2	327.5	331.28	1.15	329.98	0.76
71	aniline	training	422.61	2.1	332.4	335.59	0.96	340.48	2.43
72	aniline	training	473.53	2.1	346.8	341.84	−1.43	354.56	2.24
73	aniline	prediction	523.52	5	392.2	370.13	−5.63	376.53	−3.99
74	aniline	validation	574.26	10.1	519.9	485.41	−6.63	484.48	−6.81
75	aniline	training	523.6	30.2	348.6	335.3	−3.81	327.79	−5.97
76	aniline	validation	573.58	29.9	384.9	352.4	−8.44	350.81	−8.86
77	*o*-toluidne	training	303.55	0.1	410.5	411.29	0.19	399.79	−2.61
78	*o*-toluidne	prediction	372.23	2.2	405.1	407.32	0.55	404.27	−0.20
79	*o*-toluidne	training	422.61	2.1	402.6	400.03	−0.64	409.15	1.63
80	*o*-toluidne	training	473.53	2.1	431.4	406.3	−5.82	415.43	−3.70
81	*o*-toluidne	validation	523.53	5	482.1	441.88	−8.34	447.07	−7.27
82	*o*-toluidne	training	573.69	10.1	638.6	640.67	0.32	663.27	3.86
83	*o*-toluidne	training	523.59	30	418.8	399.47	−4.62	407.31	−2.74
84	*o*-toluidne	prediction	573.73	29.8	472.7	451.19	−4.55	445.92	−5.66
85	*m*-toluidne	training	303.55	0.1	406.2	405.8	−0.1	396.77	−2.32
86	*m*-toluidne	validation	372.2	2.1	406.1	401.42	−1.15	403.40	−0.67
87	*m*-toluidne	training	422.61	2.1	393.4	396.64	0.82	407.58	3.60
88	*m*-toluidne	training	473.53	2.1	412.9	407.33	−1.35	412.28	−0.15
89	*m*-toluidne	prediction	523.53	5	472.3	441.54	−6.51	428.38	−9.30
90	*m*-toluidne	validation	574.22	10.1	629.5	626.98	−0.4	532.74	−15.37
91	*m*-toluidne	training	523.59	30	414	398.17	−3.82	403.11	−2.63
92	*m*-toluidne	validation	573.73	29.8	456.4	448.2	−1.8	424.61	−6.97
93	*p*-toluidne	training	303.55	0.1	400.2	398.07	−0.53	392.29	−1.98
94	*p*-toluidne	prediction	372.24	2.2	398.7	395.03	−0.92	400.27	0.39
95	*p*-toluidne	training	422.61	2.1	395.4	394.91	−0.12	405.61	2.58
96	*p*-toluidne	training	473.53	2.1	426.2	412.7	−3.17	412.23	−3.28
97	*p*-toluidne	validation	523.53	5	476.5	445.21	−6.57	437.80	−8.12
98	*p*-toluidne	training	573.71	10.1	639.6	597.01	−6.66	607.16	−5.07
99	*p*-toluidne	training	523.59	30	412.5	396.55	−3.87	400.60	−2.88
100	*p*-toluidne	prediction	573.75	29.9	462.3	443.84	−3.99	433.11	−6.31
101	*m*-aminophenol	training	304.47	0.1	285	285.34	0.12	307.12	7.76
102	*m*-aminophenol	validation	372.23	2.2	303.2	308.51	1.75	308.08	1.61
103	*m*-aminophenol	training	422.61	2.1	328.7	328.88	0.05	309.60	−5.81
104	*m*-aminophenol	training	473.53	2.1	332	331.87	−0.04	312.64	−5.83
105	*m*-aminophenol	prediction	523.53	5	342.6	295.57	−13.73	320.64	−6.41
106	*m*-aminophenol	training	574.35	10.1	340.5	340.54	0.01	365.57	7.36
107	*m*-aminophenol	prediction	523.6	30.2	325.4	286.02	−12.1	308.36	−5.24
108	*m*-aminophenol	validation	573.55	29.9	323.8	307.03	−5.18	316.05	−2.39
109	*o*-diaminobenzene	training	303.55	0.1	386.5	386.27	−0.06	388.51	0.52
110	*o*-diaminobenzene	prediction	372.25	2.2	395.8	390.46	−1.35	397.57	0.45
111	*o*-diaminobenzene	training	422.65	2.1	417.9	406.06	−2.83	403.67	−3.41
112	*o*-diaminobenzene	training	473.55	2.1	433.9	437.4	0.81	410.13	−5.48
113	*o*-diaminobenzene	validation	523.55	5	465	461.28	−0.8	430.10	−7.50
114	*o*-diaminobenzene	prediction	574.55	10.2	525.3	517.23	−1.54	560.65	6.73
115	*o*-diaminobenzene	training	523.65	30.2	426.1	392.31	−7.93	396.62	−6.92
116	*o*-diaminobenzene	prediction	573.75	30	443.2	431.98	−2.53	423.31	−4.49

[a]The calculated values of $C^o_{p,2}$ and IPD using the ANN model with architecture 3-9-1. [b]The calculated values of $C^o_{p,2}$ and IPD using the ANN model with architecture 6-3-1.

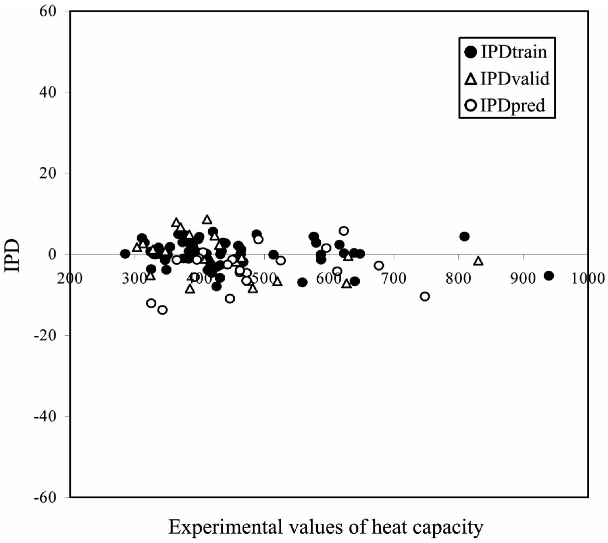

Figure 3. Plot of the IPD (individual percent deviation) for calculated values of the heat capacity from the ANN model with architecture 3-6-1 *versus* the experimental values of it for training, validation and prediction sets.

Table 2. Statistical parameters obtained by the ANN model with architecture 3-9-1 for total, training, validation and prediction sets[a]

Type of data set	N	R	MPD	RMSE	F
Total	116	0.9859	3.017	19.642	3950.35
Training	74	0.9915	2.163	14.608	4155.31
Validation	21	0.9841	4.262	22.975	584.37
Prediction	21	0.9761	4.755	29.179	383.29

[a] N is number of data set; R is the correlation coefficient between calculated and the experimental values of the partial molar heat capacity at infinite dilution; MPD is mean percent deviation; RMSE is root mean square error and F is the statistical F-value.

control the overtraining of the network during the training procedure, the values of RMSET and RMSEV were calculated and recorded to monitor the extent of the learning in various iterations. Results obtained show that overfitting does not exist for this ANN and training is stop after 80000 iterations (Figure 5).

For evaluation predictive power of the generated ANN, an optimized network was applied for prediction the heat capacity of different aqueous solutions at various temperatures and pressures in the prediction set.

Values of the partial molar heat capacity for different aqueous solutions of various polar aromatic compounds along with the calculated and IPD values at various temperatures and pressures for training, validation and prediction sets have been shown in Table 1.

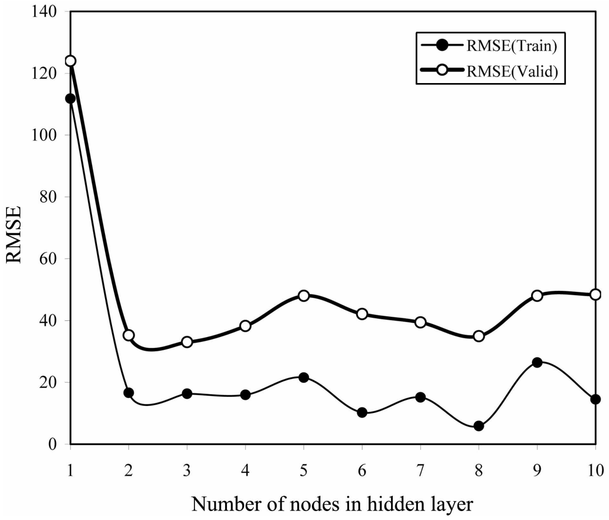

Figure 4. Plot of RMSE for training and validation sets *versus* the number of nodes in hidden layer.

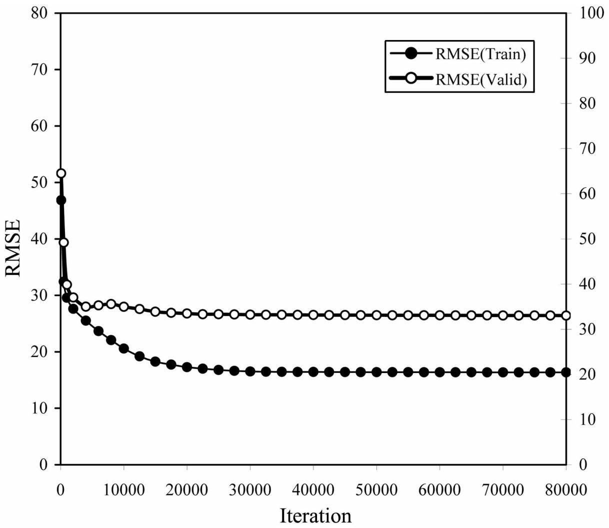

Figure 5. Plot of RMSE for training and validation sets (for the ANN model with architecture 6-3-1) *versus* the number of iterations.

As can be seen the calculated values of the heat capacity are in good agreement with those of the experimental values. The correlation equation for all of the calculated values of the heat capacity from the ANN model and the experimental

Table 3. Theoretical descriptors, symbols and coefficients in the MLR model

Name of descriptor	Symbol	Coefficient
Complementary information content (neighborhood symmetry of 0-order)	CIC0	382.718
Geary autocorrelation-lag3/weighted by atomic masses	GATS3m	−16751.82
Radial distribution function-5.0/weighted by atomic masses	RDF050m	−12.754
3D-MoRSF-signal 08/weighted by atomic polarizabilities	Mor08p	−181.68
Constant		−433.178

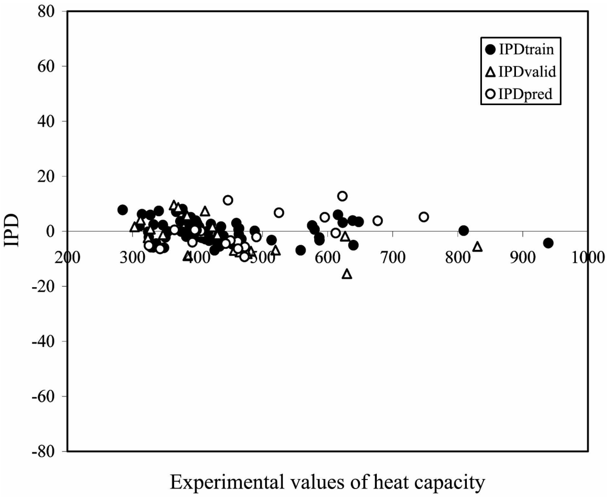

Figure 6. Plot of the IPD (individual percent deviation) for calculated values of the heat capacity from the ANN model with architecture 6-3-1*versus* the experimental values of it for training, validation and prediction sets.

Table 4. Statistical parameters obtained by the ANN model with architecture 6-3-1 for total, training, validation and prediction sets^a

Type of data set	N	R	MPD	RMSE	F
Total	116	0.9800	3.819	23.015	2762.71
Training	74	0.9893	3.141	16.348	3305.60
Validation	21	0.9745	5.386	33.028	358.61
Prediction	21	0.9815	4.642	29.885	498.20

^{a}N is number of data set; R is the correlation coefficient between calculated and the experimental values of the partial molar heat capacity at infinite dilution; MPD is mean percent deviation; RMSE is root mean square error and F is the statistical F-value.

values is as follows:

$$C^o_{p,2}\,(\text{cal}) - 0.9760\ C^o_{p,2}\,(\text{exp}) + 8.894 \qquad (7)$$

$$N = 116;\ R = 0.9800;\ MPD = 3.819;$$
$$RMSE = 23.015;\ F = 2762.71$$

Similarly, the correlation of $C^o_{p,2}$ (cal) values *versus* $C^o_{p,2}$ (exp) in prediction set gives equation (8):

$$C^o_{p,2}\,(\text{cal}) = 1.154\ C^o_{p,2}\,(\text{exp}) - 71.783 \qquad (8)$$

$$N - 21;\ R - 0.9815;\ MPD - 4.642;$$
$$RMSE - 29.885;\ F - 498.20$$

The results demonstrate that the MPD value for $C^o_{p,2}$ values in the prediction set is 4.642.

Plot of IPD for Cp values in prediction set *versus* the experimental values of it has been illustrated in Figure 6. As can be seen the model did not show proportional and systematic error, because the propagation of errors in both sides of zero are random.

The correlation coefficient (R), RMSE, MPD and statistical F-value of the model for total, training, validation and prediction sets show potential of the ANN model for prediction the heat capacity of the aqueous solutions at various temperatures and pressures (Table 4).

As a result, it was found that the properly selected and trained neural networks could fairly represent the dependence of the heat capacity of the aqueous solutions on theoretical descriptors, temperatures and pressure.

Conclusions

Two types of inputs have been applied for prediction partial molar heat capacity of aqueous solutions at infinite dilution for various polar aromatic compounds (including phenol, *o*-cresol, *m*-cresol, *p*-cresol, aniline, *o*-toluidine, *m*-toluidine, *p*-toluidine, *m*-aminophenol, *p*-aminophenol and *o*-diaminobenzene) over wide range of temperatures (303.55 -623.20 K) and pressures (0.1-30.2 MPa) using artificial neural network models. In these models macroscopic and microscopic properties of the compounds along with temperature and pressure have been used as inputs and their output is the partial molar heat capacity. The MPD values of the models for prediction set are 4.755 and 4.642, respectively. Then the optimized neural network could simulate the complicated nonlinear relationship between the partial molar heat capacity for various polar aromatic compounds on the heat capacity in T = 303.55 K and P = 0.1 MPa (or theoretical molecular descriptors), temperature and pressure. As a result ANNs can be used to predict the heat capacity at higher temperatures and pressures using minimum number of experiments.

Acknowledgements. The Authors wish to acknowledge the vice presidency of research, university of Mohaghegh Ardebili, for financial support of this work.

References

1. Censky, M.; Hnedkovsky, L.; Majer, V. *J. Chem. Thermodyn.* **2004**, *37*, 203.
2. Censky, M.; Hnedkovsky, L.; Majer, V. *J. Chem. Thermodyn.* **2005**, *37*, 225.
3. Katritzky, A. R.; Karelson, M.; Lobanov, V. S. *Pure Appl. Chem.* **1997**, *69*, 245.
4. McClelland, H. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967.
5. Habibi-Yangjeh, A. *Indian J. Chem. B* **2004**, *43*, 1504.
6. Cronce, D. T.; Famini, G. R.; Soto, J. A. D.; Wilson, L. Y. *J. Chem. Soc. Perkin Trans. 2* **1998**, 1293.
7. Engberts, J. B. F. N.; Famini, G. R.; Perjessy, A.; Wilson, L. Y. *J. Phys. Org. Chem.* **1998**, *11*, 261.
8. Nikolic, S.; Milicevic, A.; Trinajstic, N.; Juric, A. *Molecules* **2004**, *9*, 1208.
9. Devillers, J. *SAR and QSAR Environ. Res.* **2004**, *15*, 501.
10. Karelson, M.; Lobanov, V. S. *Chem. Rev.* **1996**, *96*, 1027.
11. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
12. Kramer, R. *Chemometric Techniques for Quantitative Analysis*; Marcel Dekker: New York, 1998.
13. Kuzmanovski, I.; Aleksovska, S. *Chemom. Intell. Lab. Syst.* **2003**, *67*, 167.
14. Barros, A. S.; Rutledge, D. N. *Chemomet. Intell. Lab. Syst.* **1998**, *40*, 65.
15. Garkani-Nejad, Z.; Karlovits, M.; Demuth, W.; Stimpfl, T.; Vycudilik, W.; Jalali-Heravi, M.; Varmuza, K. *J. Chromatogr. A*

2004. *1028*. 287.

16. Bose. N. K.; Liang. P. *Neural Network Fundamentals*; McGraw-Hill: New York. 1996.
17. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinhein, 1999.
18. Agatonovic-Kustrin. S.; Beresford, R. *J. Pharm. Biomed. Anal.* **2000**. *22*. 717.
19. Despagne. F.; Massart. D. L. *Analyst* **1998**. *123*. 157R.
20. Borosy. A. P.; Balogh. B.; Matyus. P. *J. Mol. Struc.(Theochem)* **2005**, *729*, 169.
21. Bunz, A. P.; Braun. B.; Janowsky, R. *Fluid Phase Equilib.* **1999**. *158*, 367.
22. Homer. J.; Generalis. S. C.; Robson. J. H. *Phys. Chem. Chem. Phys.* **1999**. *1*. 4075.
23. Goll. E. S.; Jurs. P. C. *J. Chem. Inf. Comp. Sci.* **1999**. *39*. 974.
24. Vendrame, R.; Braga. R. S.; Takahata, Y.; Galvao. D. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*. 1094.
25. Gaspelin. M.; Tusar. L.; Smid-Korbar. J.; Zupan. J.; Kristl. J. *Int. J. Pharm.* **2000**. *196*. 37.
26. Gini. G.; Cracium. M. V.; Konig. C.; Benfenati. E. *J. Chem. Inf. Comput. Sci.* **2004**. *44*. 1897.
27. Urata, S.; Takada. A.; Uchimaru, T.; Chandra. A. K.; Sekiya. A. *J. Fluorine Chem.* **2002**, *116*. 163.
28. Koziol. J. *Internet Electron. J. Mol. Des.* **2002**, *1*, 80.
29. Wegner. J. K.; Zell. A. *J. Chem. Inf. Comput. Sci.* **2003**. *43*. 1077.
30. Valkova. I.; Vracko. M.; Basak. S. C. *Anal. Chim. Acta* **2004**. *509*. 179.
31. Sebastiao. R. C. O.; Braga, J. P.; Yoshida, M. I. *Thermochimica Acta* **2004**. *412*, 107.
32. Jalali-Heravi. M.; Masoum. S.; Shahbazikhah. P. *J. Magn. Reson.* **2004**. *171*. 176.
33. Habibi-Yangjeh. A.; Nooshyar. M. *Bull. Korean Chem. Soc.* **2005**. *26*. 139.
34. Habibi-Yangjeh, A.; Nooshyar. M. *Physics and Chemistry of Liquids* **2005**. *43*. 239.
35. Habibi-Yangjeh. A.; Danandeh-Jenagharad. M.; Nooshyar. M. *Bull. Korean Chem. Soc.* **2005**. *26*. 2007.
36. Habibi-Yangjeh. A.; Danandeh-Jenagharad. M.; Nooshyar. M. *J. Mol. Model.* **2006**, *12*. 338.
37. Habibi-Yangjeh, A.; Danandeh-Jenagharad, M. *Indian J. Chem. B* **2007**. *46*, 478.
38. Habibi-Yangjeh. A. *Physics and Chemistry of Liquids* (in press).
39. Guha. R.; Jurs. P. C. *J. Chem. Inf. Comput. Sci.* **2004**. *44*. 1440.
40. Mosier. P. D.; Counterman. A. E.; Jurs. P. C.; Clemmer. D. E. *Anal. Chem.* **2002**, *74*. 1360.
41. Guha. R.; Jurs. P. C. *J. Chem. Inf. Model.* **2005**, *45*. 800.
42. HyperChem. Release 7.0 for Windows. *Molecular Modeling System*. Hypercube Inc: 2002.
43. Todeschini. R.; Consonni. V.; Pavan. M. *Dragon Software Version 2.1*; 2002.
44. Hemmateenejad, B.; Akhond. M.; Miri, R.; Shamsipur. M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1328.
45. Hemmateenejad, M.; Safarpour. M. A.; Miri, R.; Nesari, N. *J. Chem. Inf. Model.* **2005**. *45*. 190.
46. *SPSS for windows*. Release 10.0; SPSS Inc.; 1989-1999.
47. Demuth. H.; Beale. M. *Neural Network Toolbox*; Mathworks: Natick, MA, 2000.
48. Matlab 6.5. *Mathworks* 1984-2002.