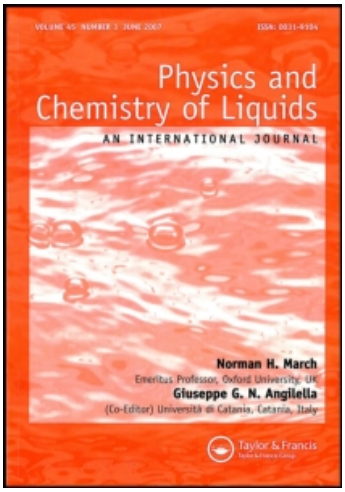

Physics and Chemistry of Liquids
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713646857

Prediction of normalized polarity parameter in binary mixed solvent systems using artificial neural networks

Aziz Habibi-Yangjeh[a]; Mahdi Nooshyar[a]
[a] Department of Chemistry, Faculty of Science, University of Mohaghegh Ardebili, Ardebil, Iran

To cite this Article Habibi-Yangjeh, Aziz and Nooshyar, Mahdi(2005) 'Prediction of normalized polarity parameter in binary mixed solvent systems using artificial neural networks', Physics and Chemistry of Liquids, 43: 3, 239 — 247
To link to this Article: DOI: 10.1080/00319100500061233
URL: http://dx.doi.org/10.1080/00319100500061233

Physics and Chemistry of Liquids
Vol. 43, No. 3, June 2005, 239–247

Taylor & Francis
Taylor & Francis Group

Prediction of normalized polarity parameter in binary mixed solvent systems using artificial neural networks

AZIZ HABIBI-YANGJEH* and MAHDI NOOSHYAR

Department of Chemistry, Faculty of Science, University of Mohaghegh Ardebili, P.O. Box 179, Ardebil, Iran

(*Received 8 November 2004*)

Artificial neural networks (ANNs) were successfully developed for the modeling and prediction of normalized polarity parameter (E_T^N) in binary mixed solvent systems at various temperatures and compositions. A three-layered feed forward ANN with back-propagation of error was generated using five parameters as inputs and its output is E_T^N. It was found that a properly selected and trained neural network could fairly represent the dependence of normalized polarity parameter on temperature and composition. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the E_T^N values in the prediction set, which were not used in the modeling procedure. Correlation coefficient (R) and root mean square error for prediction set are 0.9961 and 0.01187, respectively. The maximum value of IPD (individual percent deviation) for E_T^N values in the prediction set is 5.116%. The results show non-linear dependence of E_T^N to temperature and composition in binary mixed solvent systems is significant.

Keywords: Artificial neural networks; Solvatochromic parameters; Normalized polarity parameter; Binary mixed solvent systems

1. Introduction

The energetic level of molecules may be modified by interactions with surrounding molecules and it may be difficult to relate chemical properties to molecular structures [1–7]. The solvent effects play a key role in many chemical and physical processes in solutions. The strong influence of solvent on chemical and physical processes (reaction rates, selectivity, chemical equilibria, position and intensity of spectral absorption bands and liquid chromatographic separations) has been well established [1]. The study of solute–solvent interactions in binary mixed solvent systems is more complex than in pure solvents. On the one hand, the solute can be preferentially solvated by any of the solvents present in the mixture. On the other hand, solvent–solvent interactions can strongly affect solute–solvent interactions [8–14]. The use of solvatochromic

*Corresponding author. Email: habibiyangjeh@yahoo.com & ahabibi@uma.ac.ir

ISSN 0031-9104 print: ISSN 1029-0451 online
DOI: 10.1080/00319100500061233

indicators is a suitable method for studying solute–solvent interactions, since the transition energy of the indicators depends on the solvation's sphere composition and properties [1,15]. The solvatochromic parameter $E_T(30)$ is calculated from the maxima of absorbance of the betaine dye, expressed in wavenumber as kK ($1\,\mathrm{kK} = 1000\,\mathrm{cm}^{-1}$). It was proposed by Reichardt for measuring empirically the polarity of solvents [15]. Solvatochromic parameters are demonstrated to be successful in correlating a wide range of chemical and physical properties involving solute–solvent interactions as well as biological activities of compounds [1]. Normalized polarity parameter (E_T^N) in reference to tetramethylsilane (TMS) and water is recommended instead of $E_T(30)$ [15]:

$$E_T^N = \frac{E_T(30)_{\mathrm{Solvent}} - E_T(30)_{\mathrm{TMS}}}{E_T(30)_{\mathrm{Water}} - E_T(30)_{\mathrm{TMS}}} \tag{1}$$

Determination of E_T^N in all possible solvent compositions and temperatures in binary mixed solvent systems by experiments is hardly possible, because an infinite number of different solvent compositions can be prepared for particular binary solvent systems and also it is time consuming, since there is an infinite number of solvent compositions and temperatures that the values of E_T^N are known. For these reasons the prediction of solvatochromic parameters in binary mixed solvent systems based on a minimum number of experiments provides a useful computational tool. There are correlations between normalized polarity parameters of pure solvents and their molecular structures [16,17], but there is not any method for predicting E_T^N at various temperatures and compositions in binary mixed solvent systems.

Artificial neural networks (ANNs) have become popular due to their success where complex non-linear relationships exist amongst data [18–20]. ANNs are biologically inspired computer programs designed to simulate the way in which the human brain processes information [19,20]. ANNs gather their knowledge by detecting the patterns and relationships in data and learn (or are trained) through experience, not from programming. There are many types of neural networks designed by now and new ones are invented every week. The behavior of a neural network is determined by transfer functions of its neurons, by learning rule and by the architecture itself [19,20]. An ANN is formed from artificial neuron or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers. The first layer is termed the input layer and the last layer is the output layer. The layers of neurons between the input and output layers are called hidden layers. The wide applicability of ANNs stems from their flexibility and ability to model non-linear systems without prior knowledge of an empirical model. Neural networks do not need an explicit formulation of the mathematical or physical relationships of the handled problem. These give ANNs an advantage over traditional fitting methods for some chemical application. For these reasons, in recent years, ANNs have been used for 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 and prediction of various physico-chemical properties of compounds [21–32].

In this work, for inspection of non-linear relations of normalized polarity parameter with temperature and composition in binary mixed solvent systems, an ANN model, for the first time, was generated for the prediction of E_T^N values in various temperatures and compositions and the results were compared with the experimental values.

2. Theory

A detailed description of theory behind a neural network has been adequately described by different researchers [18–20]. There are many types of neural network architectures, but the type that has been most useful for QSAR/QSPR studies is the multilayer feed-forward network with back-propagation (BP) learning rule [20]. The numbers of neurons in the input and output layers are defined by the system's properties. The number of neurons in the hidden layer could be considered as an adjustable parameter, which should be optimized. The input layer receives the experimental or theoretical information. The output layer produces the calculated values of dependent variable. The use of ANNs consists of two steps: ''training'' and ''prediction''. In the training phase, the optimum structure, weight coefficients and biases are searched for. These parameters are found from a training data set. After the training phase, the trained network can be used to predict (or calculate) the outputs from a set of inputs. ANNs allow one to estimate relationships between input variables and one or several output dependent variables. Information from inputs is fed forward through the network to optimize the weights between neurons. Optimization of the weights is made by BP of the error during the training or learning phase. The ANN reads the input and target values in the training data set and changes the values of the weighted links to reduce the difference between the calculated output and target values. The error between output and target values is minimized across many training cycles until the network reaches a specified level of accuracy. If a network is left to train for too long, however, it will overtrain and will lose the ability to generalize [31].

3. Experimental

3.1. *Data set*

A reliable database is critically important for the training of ANNs. Fourteen binary solvent systems including mixtures of methanol + propan-1-ol, propan-1-ol + acetonitrile, dimethylsulfoxide + 2-methylpropan-2-ol, dimethylsulfoxide + propan-2-ol, dimethylsulfoxide + methanol, dimethylsulfoxide + H_2O, acetonitrile + 2-methylpropan-2-ol, acetonitrile + propan-2-ol, acetonitrile + methanol, acetonitrile + H_2O, nitromethane + 2-methylpropan-2-ol, nitromethane + propan-2-ol, nitromethane + methanol, nitromethane + H_2O, that the E_T^N values at various temperatures and compositions are available in literature have been used [33,34]. The data set was randomly divided into three groups: a training set, a validation set and a prediction set consisting of 190, 54 and 54 data, respectively. The training and validation sets were used for the model generation and the prediction set was used for the evaluation of the generated model, because a prediction set is a better estimator of the ANN generalization ability than a monitoring (validation) set [35].

3.2. *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. The inputs

of ANNs are normalized polarity parameters and mole fractions of constituents of binary mixed solvent system and temperature. A three-layer network with a sigmoidal transfer function was 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 Levenberg–Marquardt algorithms for BP of error, which, although requiring far more extensive computer memory, is significantly faster than other algorithms based on gradient descent [36]. For evaluation of the prediction power of the network, the trained ANN was used to predict E_T^N values of the data set included in the prediction set. The performances of training, validation and prediction of ANNs are evaluated by the mean percentage deviation (MPD) and root-mean square error (RMSE), which are defined as follows:

$$\mathrm{MPD} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(P_i^{\exp} - P_i^{\mathrm{cal}})}{P_i^{\exp}}\right| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N}\frac{(P_i^{\exp} - P_i^{\mathrm{cal}})^2}{N}} \tag{3}$$

where $P_i^{\exp}$ and P_i^{cal} are experimental and calculated values of E_T^N with the ANN model and N denotes the number of data points.

Individual percent deviation (IPD) is defined as follows:

$$\mathrm{IPD} = 100 \times \left|\frac{P_i^{\exp} - P_i^{\mathrm{cal}}}{P_i^{\exp}}\right| \tag{4}$$

The processing of the data was carried out on an Intel Pentium III processor, 800 MHz PC with 256 MB of RAM in Windows XP environment using Matlab 6.5 [37]. The neural networks were implemented using Neural Network Toolbox Ver. 4.0 for Matlab [38].

4. Results and discussion

There are no rigorously theoretical principles for choosing the proper network topology, so different structures were tested in order to obtain the optimal hidden neurons and training cycles [39]. Before training the network, the number of nodes in the hidden layer was 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 fifteen). The root mean squared error of training (RMSET) and validation (RMSEV) sets were plotted *versus* the number of iterations for a different number of neurons at the hidden layer and the minimum value of RMSEV was recorded as the optimum value. The plot of RMSET and RMSEV *versus* the number of nodes in the hidden layer has been shown in figure 1. It is clear that ten nodes in the hidden layer is the optimum value.

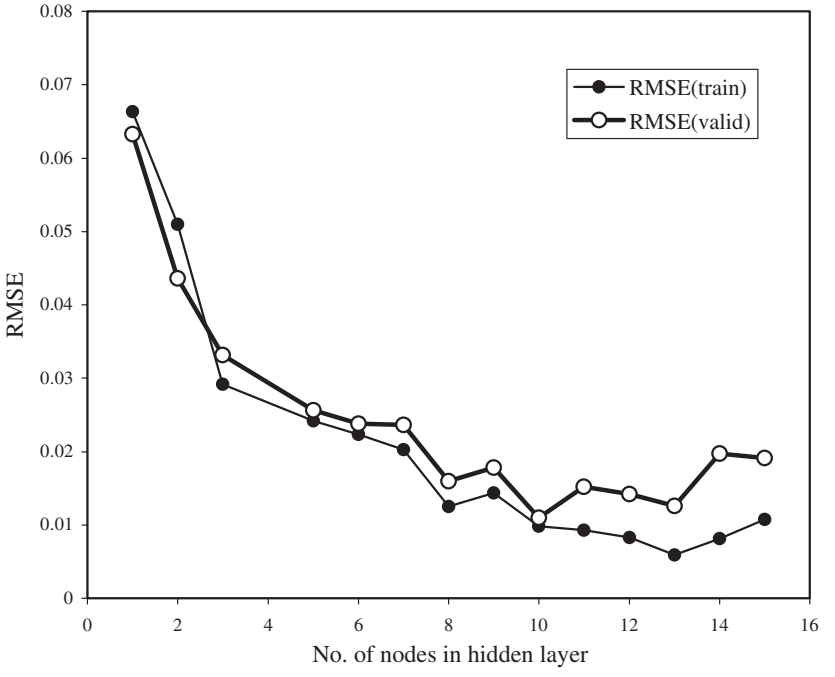

Figure 1. Plot of RMSE for training and validation sets *versus* the number of nodes in hidden layer.

This network consists of five inputs (including normalized polarity parameters and mole fractions of constituents of mixture and temperature) and one output for E_T^N. Then an ANN with architecture 5-10-1 was generated. It is noteworthy that training of the network was stopped when the RMSEV started to increase i.e. when overtraining begins. The overtraining causes the ANN to lose its prediction power [31]. 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. Results obtained showed that after 1275 iterations the value of RMSEV started to increase and overfitting began.

The generated ANN was then trained using the training set for the optimization of the weights and biases. For the evaluation of the predictive power of the generated ANN, an optimized network was applied for prediction of the E_T^N values in various binary mixed solvent systems at different temperatures and compositions in the prediction set, which were not used in the modeling procedure. The calculated values of the E_T^N at various temperatures and compositions for training, validation and prediction sets using the ANN model have been plotted *versus* the experimental values of it in figure 2.

Figure 3 demonstrates the plot of the calculated values of E_T^N at various temperatures and compositions *versus* the experimental values of it for the prediction set.

As expected, the calculated values of E_T^N are in good agreement with those of the experimental values. The correlation equation for all of the calculated values of E_T^N

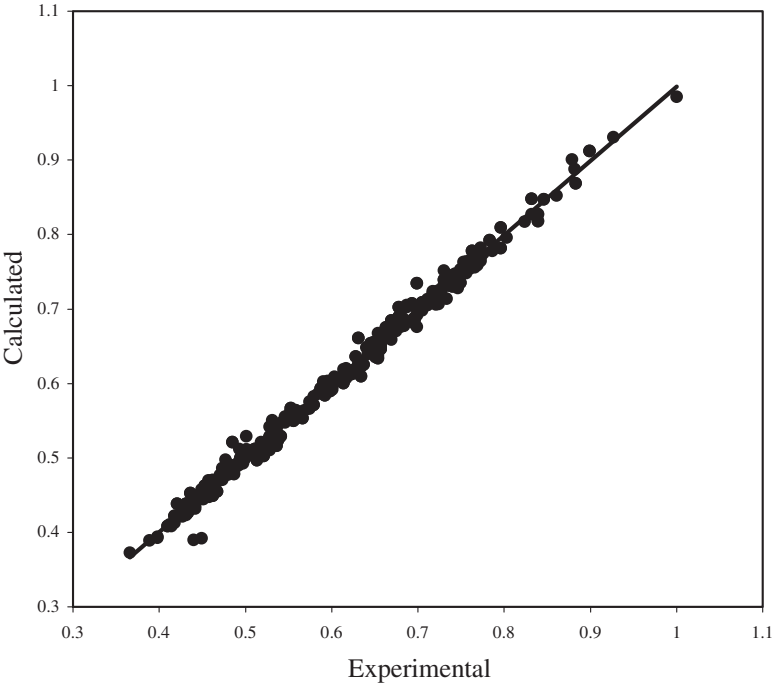

Figure 2. Plot of the calculated values of E_T^N from the ANN model *versus* the experimental values of it for total set of data.

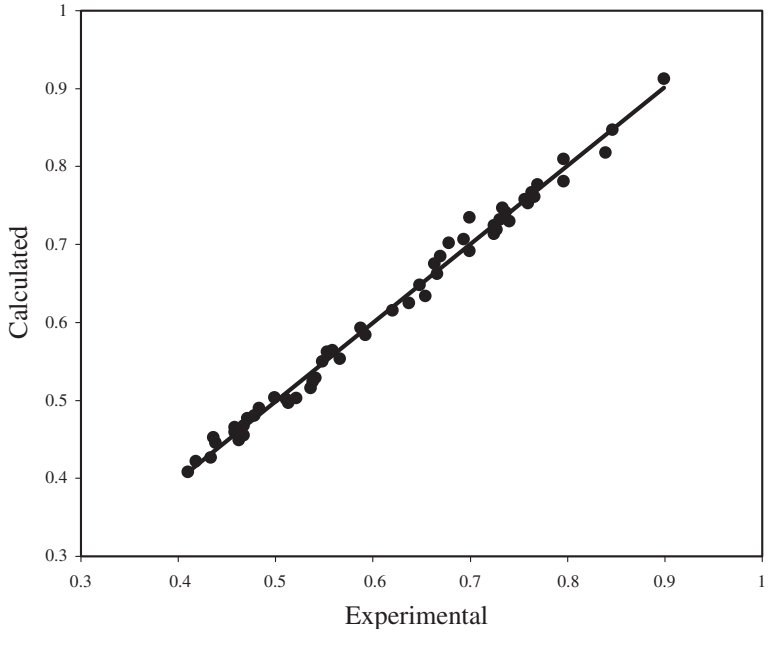

Figure 3. Plot of the calculated values of E_T^N from the ANN model *versus* the experimental values of it for prediction sets.

from the ANN model and the experimental values is as follows:

$$E_T^N(\text{cal}) = 0.99650\, E_T^N(\text{exp}) + 0.00195$$
$$R^2 = 0.9930; \quad \text{MPD} = 1.5917; \quad \text{RMSE} = 0.01046; \quad F = 42558.43 \tag{5}$$

Similarly, the correlation of E_T^N(cal) values *versus* E_T^N(exp) in the prediction set gives equation (6):

$$E_T^N(\text{cal}) = 1.00864\, E_T^N(\text{exp}) - 0.00553$$
$$R^2 = 0.9922; \quad \text{MPD} = 2.2573; \quad \text{RMSE} = 0.01187; \quad F = 6604.16 \tag{6}$$

The plot of IPD for E_T^N values in the prediction set *versus* the experimental values of it has been illustrated in figure 4. The results demonstrate that the maximum value of IPD for E_T^N values in the prediction set is 5.116%.

As can be seen, the model did not show proportional and systematic error, because the slope ($a = 1.00864$) and intercept ($b = 0.00553$) of the correlation equation are not significantly different from unity and zero, respectively and the propagation of errors in both sides of zero is random (figure 4).

Table 1 compares the results obtained using the ANN model. The correlation coefficient (R), RMSE, MPD and F-value of the model for total, training, validation and prediction sets show the potential of the ANN model for prediction of E_T^N values in binary mixed solvent systems at various temperatures and compositions.

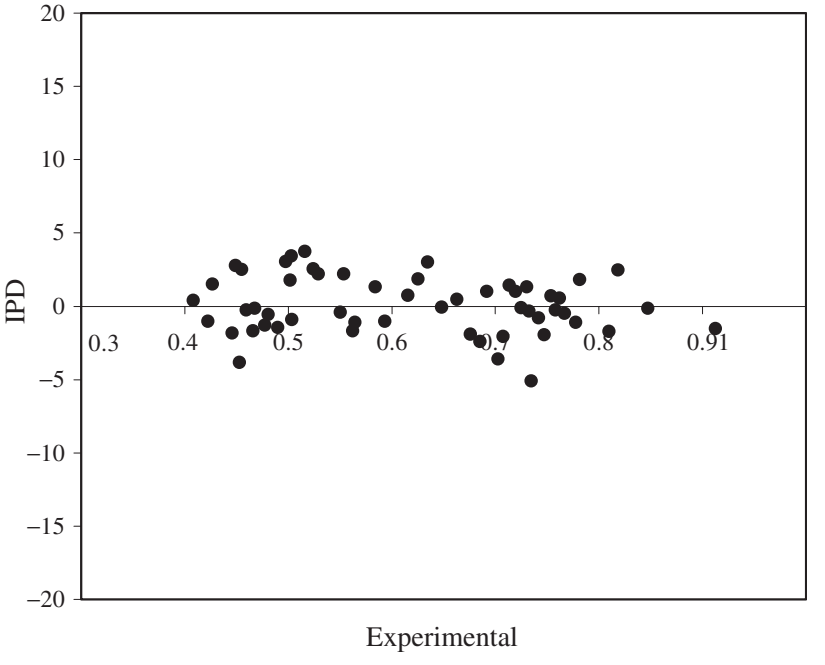

Figure 4. Plot of the IPD (individual percent deviation) for calculated values of E_T^N from the ANN model *versus* the experimental values of it for prediction sets.

Table 1. Statistical parameters obtained by ANN model for total, training, validation and prediction sets.[a]

Type of data set	N	R	MPD	RMSE	F
Total	298	0.9965	1.5917	0.01046	42558.43
Training	190	0.9969	1.6574	0.00991	30181.16
Validation	54	0.9958	1.3838	0.01103	6125.29
Prediction	54	0.9961	2.2573	0.01187	6604.16

[a]N is number of data set; R is the correlation coefficient between calculated and the experimental values of E_T^N; MPD is mean percent deviation; RMSE is root mean square error and F is the statistical F-value.

As a result, it was found that a properly selected and trained neural network could usefully represent the dependence of the normalized polarity parameter on temperatures and composition in binary mixed solvent systems. Then the optimized neural network could simulate the complicated non-linear relationship between E_T^N values of binary mixed solvent systems and E_T^N of pure solvents, mole fractions of solvents and temperature.

References

[1] C. Reichardt, *Solvents and Solvent Effects in Organic Chemistry*, 3rd Edn, Chap. 4–7, VCH (2003).
[2] Y. Marcus, *J. Chem. Soc. Perkin Trans.*, **2**, 1015 (1994).
[3] C. Cativiela, J.I. Garcia, J. Gil, R.M. Martinez, J.A. Mayoral, L. Salvatella, J.S. Urieta, A.M. Mainer, M.H. Abraham, *J. Chem. Soc. Perkin Trans.*, **2**, 653 (1997).
[4] M.R. Gholami, A. Habibi-Yangjeh, *Int. J. Chem. Kinet*, **32**, 431 (2000).
[5] M.R. Gholami, A. Habibi-Yangjeh, *J. Phys. Org. Chem.*, **13**, 468 (2000).
[6] M.R. Gholami, A. Habibi-Yangjeh, *Int. J. Chem. Kinet*, **33**, 118 (2001).
[7] A. Habibi-Yangjeh, M.R. Gholami, R. Mostaghim, *J. Phys. Org. Chem.*, **14**, 884 (2001).
[8] R.D. Shwierczynski, K.A. Connors, *J. Chem. Soc. Perkin Trans.*, **2**, 467 (1994).
[9] E. Bosch, F. Rived, M. Roses, *J. Chem. Soc. Perkin Trans.*, **2**, 2177 (1996).
[10] E. Bosch, M. Roses, K. Herodes, I. Koppel, I. Leito, I. Koppel, V. Taal, *J. Phys. Org. Chem.*, **9**, 403 (1996).
[11] C. Rafols, M. Roses, E. Bosch, *J. Chem. Soc. Perkin Trans.*, **2**, 243 (1997).
[12] M. Roses, U. Buhvestov, C. Rafols, F. Rived, E. Bosch, *J. Chem. Soc. Perkin Trans.*, **2**, 1341 (1997).
[13] U. Buhvestov, F. Rived, C. Rafols, E. Bosch, M. Roses, *J. Phys. Org. Chem.*, **11**, 185 (1998).
[14] R.E. Leitao, F. Martins, M.C. Ventura, N. Nunes, *J. Phys. Org. Chem.*, **15**, 623 (2002).
[15] C. Reichardt, *Chem. Rev.*, **94**, 2319 (1994).
[16] A.R. Katritzky, T. Tamm, Y. Wang, S. Sild, M. Karelson, *J. Chem. Inf. Comput. Sci.*, **39**, 684 (1999).
[17] G.R. Famini, L.Y. Wilson, *J. Phys. Org. Chem.*, **12**, 645 (1999).
[18] J.V. Turner, D.J. Maddalena, D.J. Cutler, *Int. J. Pharm.*, **270**, 209 (2004).
[19] N.K. Bose, P. Liang, *Neural Network Fundamentals*, McGraw-Hill (1996).
[20] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinhein (1999).
[21] S.L. Anker, P.C. Jurs, *Anal. Chem.*, **64**, 1157 (1992).
[22] W.L. Xing, X.W. He, *Anal. Chim. Acta*, **349**, 283 (1997).
[23] A.P. Bunz, B. Braun, R. Janowsky, *Fluid Phase Equilib.*, **158**, 367 (1999).
[24] J. Homer, S.C. Generalis, J.H. Robson, *Phys. Chem. Chem. Phys*, **1**, 4075 (1999).
[25] E.S. Goll, P.C. Jurs, *J. Chem. Inf. Comp. Sci.*, **39**, 974 (1999).
[26] R. Vendrame, R.S. Braga, Y. Takahata, D.S. Galvao, *J. Chem. Inf. Comput. Sci.*, **39**, 1094 (1999).
[27] M. Gaspelin, L. Tusar, J. Smid-Korbar, J. Zupan, J. Kristl, *Int. J. Pharm.*, **196**, 37 (2000).
[28] S. Urata, A. Takada, T. Uchimaru, A.K. Chandra, A. Sekiya, *J. Fluorine Chem.*, **116**, 163 (2002).
[29] J.K. Wegner, A. Zell, *J. Chem. Inf. Comput. Sci.*, **43**, 1077 (2003).
[30] R.C.O. Sebastiao, J.P. Braga, M.I. Yoshida, *Thermochimica Acta*, **412**, 107 (2004).
[31] M. Jalali-Heravi, S. Masoum, P. Shahbazikhah, *J. Magn. Reson.*, **171**, 176 (2004).
[32] I. Valkova, M. Vracko, S.C. Basak, *Anal. Chim. Acta*, **509**, 179 (2004).
[33] E. Bosch, M. Roses, *J. Phys. Org. Chem.*, **9**, 403 (1996).
[34] R.E. Leitao, F. Martins, M.C. Ventura, N. Nunes, *J. Phys. Org. Chem.*, **15**, 623 (2002).

[35] D. Svoizl, V. Kvasnicka, J. Pospichal, *Chemom. Intell. Lab. Syst.*, **39**, 43 (1997).
[36] M.T. Hagan, M. Menhaj, *IEEE Trans. Neural Networks*, **5**, 989 (1994).
[37] Matlab 6.5, *Mathworks* (1984–2002).
[38] H. Demuth, M. Beale, *Neural Network Toolbox*, Mathworks, Natick, MA (2000).
[39] R.G. Alany, S. Agatonovic-kustrin, T. Rades, L.G. Tucker, *J. Pharm. Biomed. Anal.*, **19**, 443 (1999).