Journal of Hazardous Materials 164 (2009) 182–194

Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

# Estimation of toxicity of ionic liquids in *Leukemia Rat Cell Line* and *Acetylcholinesterase* enzyme by principal component analysis, neural networks and multiple lineal regressions

José S. Torrecilla*, Julián García, Ester Rojo, Francisco Rodríguez

*Department of Chemical Engineering, Faculty of Chemistry, Complutense University of Madrid, 28040 Madrid, Spain*

ARTICLE INFO

*Article history:*
Received 27 February 2008
Received in revised form 24 July 2008
Accepted 1 August 2008
Available online 15 August 2008

*Keywords:*
Ionic liquid
Toxicity
PCA
Mathematical model

ABSTRACT

Multiple linear regression (MLR), radial basis network (RB), and multilayer perceptron (MLP) neural network (NN) models have been explored for the estimation of toxicity of ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium and quinolinium ionic liquid salts in the *Leukemia Rat Cell Line* (IPC-81) and *Acetylcholinesterase* (AChE) using only their empirical formulas (elemental composition) and molecular weights. The toxicity values were estimated by means of decadic logarithms of the half maximal effective concentration ($EC_{50}$) in μM ($\log_{10} EC_{50}$). The model's performances were analyzed by statistical parameters, analysis of residuals and central tendency and statistical dispersion tests. The MLP model estimates the $\log_{10} EC_{50}$ in IPC-81 and AchE with a mean prediction error less than 2.2 and 3.8%, respectively.

## 1. Introduction

In the recent years, ionic liquids (ILs) have been given an increasing attention in the chemical industry sector because of their attractive properties as solvents (negligible vapor pressure, high dissolving power, thermomechanical and electrochemical stability, a wide range in the liquid state, solvating properties for diverse materials, high ionic conductivity, wide electrochemical window, etc.). Due to their negligible vapor pressures, ILs are potentially considered as environmentally friendly [1]. But as ILs are soluble in water, they can accumulate in the environment, and so the determination of toxicity is required to determine the environmental risk of accidental ILs discharges.

In recent years, different studies have been carried out to determinate the toxicity of ILs. In particular, the influence of ILs on *Vibrio fischeri*, green *algae* species, *Daphnia magna*, *Lemna minor*, *Danio rerio*, *Acetylcholinesterase* and *Leukemia Rat Cell Line* have been studied [2–4].

The high number of possible combinations between cations and anions leads to an enormous number of viable ILs. Up to now, more than a million different ionic liquids have been designed [5]. Every day, thanks to the high number of industrial applications and advantages of using ILs, others ILs are appearing. But, from an environmental point of view, the number of ILs with known effects on the environment has increased at a slower pace. Moreover, concerning the risk assessment for human beings, the number of ILs with known effects is increased even more slowly. For this reason, designing a mathematical tool to estimate the toxicity of ILs is very important. With this objective, Luis et al. have designed an algorithm to estimate the aquatic toxicity of 43 imidazolium, pyridinium and pyrrolidinium ILs using a novel group contribution. The correlation coefficient between estimated values by the proposed model and their respective real values is higher than 0.9 [6]. In this line, but estimating the toxicity of non-ionic liquids compounds, investigations can be found in literature. For example, the pesticide aquatic toxicity was studied by Mazzatorta et al. using linear and non-linear regressions [7]. Zhao et al. estimated the toxicity of organic compounds by neural networks (NNs), multiple linear regression and molecular structures of these chemicals [8]. Gosav et al. estimated the toxicity of novel amphetamines using NNs and the constitutional characteristics of them [9]. NNs and principal component analysis (PCA) were applied to design expert systems to diagnose the atherosclerosis [10], to discover the correlation between urinary nucleoside profiles and tumours [11] and to screen novel therapeutic agents in toxicological studies [12]. Given that there is no reference related with the estimation of the toxicity of ionic liquids using NNs/PCA and the successful application of this combination in other fields, these algorithms have been tested here.

* Corresponding author. Tel.: +34 394 42 40; fax: +34 394 42 43.
*E-mail addresses:* jstorre@quim.ucm.es (J.S. Torrecilla), jgarcia@quim.ucm.es (J. García), erojo@quim.ucm.es (E. Rojo), frsomo@quim.ucm.es (F. Rodríguez).

0304-3894/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2008.08.022

Neural network is a mathematical algorithm which has the capability of relating the input and output variables without requiring a prior knowledge of the relationships between them. Its structure is relatively simple, with connections in parallel and sequence between neurons. This means a short computing time and a high potential of robustness and adaptive performance [13]. On the other hand, principal component analysis is a non-parametric and unsupervised technique, mathematically defined as an orthogonal lineal transformation. This technique transforms the data into a new coordinate system. Because of this, PCA is used to reduce the dimensionality of the database retaining their characteristics.

The aim of this work is to propose, design and validate different mathematical methods to estimate the toxicity of ionic liquids by means of available information. With this objective, molecular weights and empirical formulas (elemental composition) of 153 ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium and quinolinium ionic liquids have been used to quantify their biological activity in an *Acetylcholinesterase* (AChE) enzyme which is an essential part of the human nervous system and in *Leukemia Rat Cell Line* (IPC-81) [4]. The toxicity was evaluated by means of decadic logarithms of the half maximal effective concentration ($EC_{50}$) in μM ($\log_{10} EC_{50}$). Mathematicaly, these estimations were carried out by the combination of principal component analysis techniques and linear regressions or non-linear models (NNs models). To develop them, experimental data available in the literature were employed [2].

The paper is organized as follows: first the PCA and models are presented (a detailed explanation of them is given in Appendix A), then the dimensionality of the data shown in Table A1 (Appendix A) is reduced by PCA technique. And using these new data, all models are designed and tested. Then, their estimations are statistically analyzed and compared.

## 2. Material and methods

In this work, to design and optimize all the models used, a database of $\log_{10} EC_{50}$ in IPC-81 and AChE systems for 153 ionic liquids was obtained from literature (Table A1, Appendix A) [2]. The approximate confidence regions of these data were established to be about ±0.15 [2].

Four different models were studied. Two linear (multiple linear regression models with and without constant, MLR) and two non-linear (radial basis, RB, and multilayer perceptron neural networks, MLP) models have been tested. Every NN model used in this work was designed using Matlab version 7.01.24704 (R14) [14]. The MLR models, principal component analysis and statistical analyses were carried out by Statgraphics Plus version 5.1 [15].

### 2.1. Principal component analysis

Principal component analysis is a classical unsupervised technique based on linear algebra. It involves a mathematical procedure, described in Appendix A, that transforms a number of possible correlated variables into a smaller number of uncorrelated variables called principal components (PCs) [16]. The principal components are linear combinations of the original variables. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. This linear transformation has been widely used in data analysis, on exploratory tool to uncover unknown trends in the data, compression, etc.

### 2.2. Neural networks

Multilayer perceptron (MLP) and radial-basis function (RB) models have been used here. The MLP model is probably the most commonly used today. It is a feed-forward network with

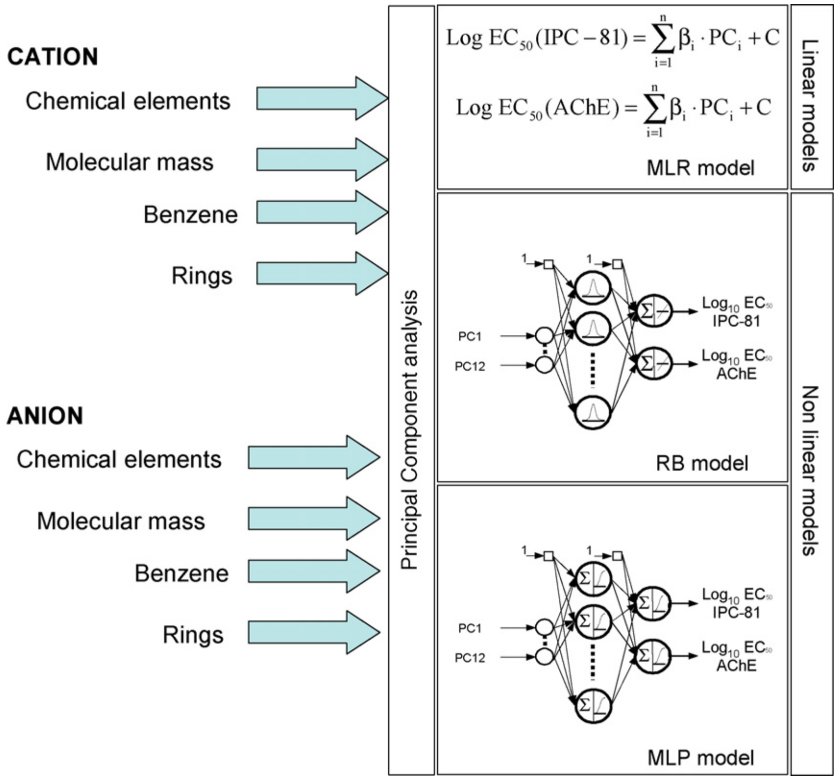
CATION
Chemical elements
Molecular mass
Benzene
Rings
ANION
Chemical elements
Molecular mass
Benzene
Rings
Principal Component analysis
$\mathrm{Log\ EC_{50}(IPC-81)} = \sum_{i=1}^{n} \beta_i \cdot PC_i + C$
$\mathrm{Log\ EC_{50}(AChE)} = \sum_{i=1}^{n} \beta_i \cdot PC_i + C$
MLR model
Linear models
1
1
PC1
PC12
$\Sigma$
$\mathrm{Log_{10}\ EC_{50}}$ IPC-81
$\mathrm{Log_{10}\ EC_{50}}$ AChE
RB model
1
1
PC1
PC12
$\Sigma$
$\mathrm{Log_{10}\ EC_{50}}$ IPC-81
$\mathrm{Log_{10}\ EC_{50}}$ AChE
MLP model
Non linear models

**Fig. 1.** Scheme of linear and non-linear models programming methodology.

**Table 1**
List of constitutional descriptors used

| Cation | Anion | |
|---|---|---|
| Number of atoms | Number of atoms | Relative number of I atoms |
| Number of C atoms | Number of C atoms | Number of S atoms |
| Relative number of C atoms | Relative number of C atoms | Relative number of S atoms |
| Number of H atoms | Number of H atoms | Number of B atoms |
| Relative number of H atoms | Relative number of H atoms | Relative number of B atoms |
| Number of N atoms | Number of N atoms | Number of Sb atoms |
| Relative number of N atoms | Relative number of N atoms | Relative number of Sb atoms |
| Number of O atoms | Number of O atoms | Number of P atoms |
| Relative number of O atoms | Relative number of O atoms | Relative number of P atoms |
| Number of P atoms | Number of Cl atoms | Number of rings |
| Relative number of P atoms | Relative number of Cl atoms | Relative number of rings |
| Number of rings | Number of Br atoms | Number of benzene rings |
| Relative number of rings | Relative number of Br atoms | Relative number of benzene rings |
| Number of benzene rings | Number of F atoms | Molecular weight |
| Relative number of benzene rings | Relative number of F atoms | |
| Molecular weight | Number of I atoms | |

a prediction horizon and supervised learning. It is characterised by layered architectures and feed-forward connections between neurons, or back connections. Weights are assigned to these connections between the neurons of one layer and the following. In order to predict with the least possible error, these values must be optimized. The MLP model is a good pattern classifier, signal filter and data compressor [17]. Specifically, it is used to model systems based on non-linear dynamics [13,18]. The other type of NN model is the RB. As can be seen in Fig. 1, the RB model shows similar topology but the transfer function and learning method are different. The design of a RB model can be viewed as a curve-fitting problem in a high-dimensional space. Accordingly, the learning is equivalent to finding a surface in a multidimensional space that provides an optimal fit [19]. Another important advantage of both NN models is that knowledge of the system to be modelled is not necessary; therefore, the NNs have enormous applicability.

As can be seen in Fig. 1, the MLP and RB models used in this work consist of two layers with connections to the outside world (an input layer where data are presented to the network and an output layer which holds the network response to given inputs) and one hidden layer. Both non-linear models use the same learning and verification samples. The characteristics of NN models, how to optimize their parameters, learning and verification samples and processes are described in Appendix A.

### 2.3. Linear models

The linear models tested in this work are considered linear in the parameters, also called statistically linear. Linear and multiple linear regressions are the most widely used and known modelling methods. They have been adapted to a broad range of situations. In a multivariate case, when there is more than one independent variable, the regression line cannot be visualized in two dimensions

**Table 2**
Principal component weight values depending on the main constitutional descriptors

| Cation | PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | PC 6 | PC 7 |
|---|---|---|---|---|---|---|---|
| Molecular weight | 0.1298 | 0.0847 | 0.1482 | 0.1701 | −0.1714 | 0.2417 | −0.8411 |
| Number of atoms | 0.1400 | −0.3838 | 0.1239 | −0.1300 | −0.0906 | 0.0053 | −0.0823 |
| Number of C atoms | 0.0314 | −0.4193 | −0.0480 | −0.1156 | −0.0861 | 0.0061 | −0.1136 |
| Relative number of C atoms | −0.2259 | −0.1704 | −0.3753 | 0.0280 | 0.0764 | 0.0658 | −0.1031 |
| Number of H atoms | 0.1989 | −0.3423 | 0.1657 | −0.1512 | −0.0632 | −0.0164 | −0.0539 |
| Relative number of H atoms | 0.3118 | −0.0155 | 0.2352 | −0.2179 | −0.0612 | −0.1763 | 0.1232 |
| Number of N atoms | −0.3493 | −0.0180 | 0.1790 | 0.1332 | 0.0913 | −0.0601 | −0.1834 |
| Relative number of N atoms | −0.3081 | 0.1417 | 0.0422 | 0.2620 | 0.2289 | 0.0555 | −0.1627 |
| Number of O atoms | 0.1218 | 0.2707 | 0.1965 | 0.0691 | −0.4113 | 0.1964 | 0.1081 |
| Relative number of O atoms | 0.1092 | 0.2794 | 0.2007 | 0.0979 | −0.3988 | 0.1922 | 0.1235 |
| Number of P atoms | 0.2138 | −0.1756 | −0.0115 | 0.3597 | 0.2311 | 0.4087 | 0.1849 |
| Relative number of P atoms | 0.2138 | −0.1756 | −0.0115 | 0.3597 | 0.2311 | 0.4087 | 0.1849 |
| Number of rings | −0.2639 | 0.0014 | −0.0439 | −0.3798 | −0.0910 | 0.4773 | 0.0689 |
| Relative number of rings | −0.2639 | 0.0014 | −0.0439 | −0.3798 | −0.0910 | 0.4773 | 0.0689 |
| Number of benzene rings | −0.1034 | −0.1279 | −0.3732 | 0.2072 | −0.4193 | −0.0809 | 0.0428 |
| Relative number of benzene rings | −0.1034 | −0.1279 | −0.3732 | 0.2072 | −0.4193 | −0.0809 | 0.0428 |

| Anion | PC 8 | PC 9 | PC 10 | PC 11 | PC 12 |
|---|---|---|---|---|---|
| Number of C atoms | 0.2580 | 0.5294 | −0.0434 | 0.0423 | 0.1315 |
| Relative number of C atoms | 0.2830 | 0.4304 | 0.0602 | −0.0498 | −0.1141 |
| Number of H atoms | 0.0912 | 0.5807 | 0.0685 | −0.0178 | 0.04761 |
| Relative number of Br atoms | −0.1526 | −0.0534 | 0.3791 | −0.0532 | 0.8426 |
| Number of F atoms | 0.2950 | −0.3185 | −0.3391 | −0.1662 | 0.0777 |
| Number of I atoms | −0.0704 | 0.0159 | −0.1053 | 0.9145 | 0.0229 |
| Number of S atoms | 0.4299 | −0.1683 | 0.03256 | 0.0400 | −0.0012 |
| Relative number of S atoms | 0.3737 | −0.1561 | 0.0881 | 0.03679 | −0.1855 |
| Number of B atoms | −0.1529 | 0.1604 | −0.5775 | −0.2988 | 0.1281 |
| Relative number of B atoms | −0.1538 | 0.0542 | −0.5961 | 0.1622 | 0.2250 |
| Molecular weight | 0.4400 | −0.0558 | −0.1009 | 0.0855 | 0.2193 |

space. In this case, a linear equation containing all those variables can be constructed in IPC-81 and AChE, Eq. (1).

$$y = \sum_{i=1}^{n} \beta_i x_i + C \qquad (1)$$

In Eq. (1), $y$, $\beta_i$, $x_i$ and $C$ represent response variable, parameters of the model, independent variables and constant of the model, respectively [20].

## 3. Result and discussion

PCA is used to reduce the dimensionality of data and with these new data two linear and two non-linear models have been designed and tested to estimate the toxicology of ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium and quinolinium ionic liquids, Table A1 (Appendix A).

### 3.1. Principal component analysis

The NN models are used to estimate the $\log_{10}$ $EC_{50}$ in IPC-81 and AChE systems in the presence of ILs studied. These models have two output neurons. As the models were designed as tool to estimate easily the toxicity of ILs, the output values were calculated using the elemental composition and molecular weight of every IL, Table 1. As can be seen, in this way, forty-six input nodes would be necessary. If a MLP model with forty-six inputs and two outputs were used, more than 256 parameters of the NN must be optimized. As the number of parameters is higher than the learning sets, this topology would not be adequate. Therefore, decreasing the number of input variables is necessary. The reduction of the number of parameters was carried out by principal component analysis technique (described in Appendix A).

The simplest and the most common method used to solve the number of principal component problem is the *eigenvalue*-one criterion also known as the Kaiser criterion [21] where the principal components with *eigenvalue* greater than 1 are selected. The PC with highest *eigenvalue* is considered as the most significant and subsequently the PCs are introduced into the calibration model one after the other until the *eigenvalue* is equal to unity [22]. As can be seen in Table 4, PCA yields seven and five PCs explaining 94.76 and 88.31% of the total variance in the cation and anion variables, respectively. The dimension of the input data was decreased from 46 independent variables to 12 principal components (7 for the cation and 5 for the anion). The principal components are calculated using Eq. (2).

$$\mathrm{PC}_j = \sum_{i=1}^{M} \mathrm{Wpc}_i^j \cdot D_i^j \qquad (2)$$

In Eq. (2), the $\mathrm{Wpc}_i^j$, $D_i^j$ and $M$ are the weight for a given principal component ($1 \leq j \leq 12$), the constitutional descriptor value and the number of descriptors used, respectively. The $\mathrm{Wpc}_i^j$ values are shown in Table 2. The three descriptors with the greatest influence on most of the principal components are shown in Table 3. The number of phosphorus atoms and molecular weight of the ILs have influence over 42% on the principal components. The effect of the molecular weight of different types of polymers on the toxicity found here has been published previously [23,24].

Given that 12 principal components are necessary, the input layer consists of 12 nodes. Then, the NN models are made up of 12 input nodes and two output neurons. Learning and verification samples were made (Table A1, Appendix A). Both samples were composed of 14 rows, one for each variable (12 principal components and 2 for $\log_{10}$ $EC_{50}$ in IPC-81 and AChE systems). The learning

**Table 3**
The three most influential constitutional descriptors on the principal components

| PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | PC 6 | PC 7 |
|---|---|---|---|---|---|---|
| Cation | | | | | | |
| Relative number of H atoms | Number of atoms | Relative number of C atoms | Number of P atoms | Number of O atoms | Number of P atoms | Molecular weight |
| Number of N atoms | Number of C atoms | Number of benzene Rings | Number of rings | Number of benzene rings | Number of rings | Number of P atoms |
| Relative number of N atoms | Number of H atoms | Relative number of benzene rings | Relative number of rings | Relative number of benzene rings | Relative number of rings | Relative number of P atoms |
| PC 8 | PC 9 | PC 10 | PC 11 | PC 12 | | |
| Anion | | | | | | |
| Number of S atoms | Number of C atoms | Relative number of Br atoms | Number of F atoms | Relative number of Br atoms | | |
| Relative number of S atoms | Relative number of C atoms | Number of B atoms | Number of I atoms | Relative number of B atoms | | |
| Molecular weight | Number of H atoms | Relative number of B atoms | Relative number of B atoms | Molecular weight | | |

**Table 4**
Main characteristics of the principal components selected

| Cation | | | Anion | | |
|---|---|---|---|---|---|
| Principal components | Eigenvalue | Explained variance (%) | Principal components | Eigenvalue | Explained variance (%) |
| PC1 | 6.0110 | 28.624 | PC8 | 4.6514 | 38.762 |
| PC2 | 5.1538 | 53.166 | PC9 | 2.2532 | 57.538 |
| PC3 | 3.1722 | 68.272 | PC10 | 1.7628 | 72.228 |
| PC4 | 1.9627 | 77.618 | PC11 | 1.0973 | 81.372 |
| PC5 | 1.6683 | 85.562 | PC12 | 1.0832 | 88.307 |
| PC6 | 1.1306 | 90.946 | | | |
| PC7 | 1.0801 | 94.760 | | | |

sample consists of 121 columns (80% of whole data). The verification sample has the same format but with the remaining 20%.

### 3.2. Multiple linear regression models

The MLR models represent the independent contributions of each principal component to the dependent variable estimation ($\log_{10} EC_{50}$). In this model, the linearity relationship between variables is assumed. Obviously, in practice, this assumption can never be confirmed, but in some cases the MLR procedures are not greatly affected by minor deviations from this assumption. In this case, the assumption reliability has been tested by calculating statistical parameters (correlation coefficient, $R^2$, standard deviation, $\sigma$, and mean prediction error, Eq. (3)). The regression coefficients and statistics of fits obtained from MLR models using verification sample are shown in Table 5. As can be seen, the data were fitted using two MLR models, Eq. (1) (with and without constant, $C$).

$$\mathrm{MPE} = \frac{1}{N}\sum_{k=1}^{N}\frac{|r_k - y_k|}{r_k}100 \tag{3}$$

In Eq. (3), $N$, $r_k$ and $y_k$ are the number of estimations, and the real and estimated values, respectively. Taking into account the mean prediction error (MPE), $R^2$ and $\sigma$ values (Table 5 and Fig. 2) calculated using the verification sample, the MLR model with constant describes the experimental data more adequately than the MLR without constant model, Eq. (1).

Although these MLR models are the best linear fit between experimental $\log_{10} EC_{50}$ and the twelve principal components, $R^2 > 0.7$ can be reached using only the four first PCs. Therefore, the constitutional descriptors of the cations present the greatest influence on the studied toxicity. This point is in agreement with literature [25,26]. In particular, the number of aromatic rings, carbon and nitrogen atoms of the cation are the most important constitutional descriptors, and these influences are also in agreement with literature [26].

### 3.3. Non-linear models

As can be seen in Table 5, a linear model with 12 or 13 parameters is not able to describe the system as non-linear models could do. Therefore, the non-linear models were tested.

**Table 6**
Correlation coefficients of real against predicted values of $\log_{10} EC_{50}$ of IPC-81 and AChE systems in ILs mixtures using RB model and verification sample

| Systems | HNN | SC | MPE (%) | $R^2$ | $\sigma$ |
|---|---|---|---|---|---|
| $\log_{10} EC_{50}$ (IPC-81) | 121 | 100 | 11.1 | 0.861 | 0.07 |
| $\log_{10} EC_{50}$ (AChE) | 121 | 1000 | 7.1 | 0.842 | 0.05 |

#### 3.3.1. RB models

Following the description presented in Appendix A, the spread constant was optimized. The main results are shown in Table 6. As can be seen, using the verification sample, the $\log_{10} EC_{50}$ in both IPC-81 and AChE systems was estimated with correlation coefficients of real vs. predicted values higher than 0.861 and 0.842, respectively. The mean prediction error values are less than 11.1 and 7.1%, respectively, Table 6.

As can be seen in Fig. 1, the RB model consists of an input, hidden and output layers. The transfer functions of hidden neurons are Gaussian types but the output layer is formed by linear functions. Because of this, the statistical results are slightly better than those determined by the application of linear models.

#### 3.3.2. MLP models

As was described in Appendix A, the parameters of the NN were optimized. The main results are shown in Table 7. In the MLP model, every neuron is based on non-linear function (sigmoid function). In this case, using the verification sample, the mean prediction error (MPE) calculated in the estimation of IPC-81 and AChE is less than 3.9 and 2.8%, respectively, Table 7. The MLP model uses less than 61 parameters, Table 7. Moreover, in most systems, estimated value/real value = 1.001, Fig. 2.

### 3.4. Models comparison

In order to guarantee the reliability of the estimations calculated by these models, the applicability domain has been evaluated selecting the compounds with cross-validated standardized residuals greater than three standard deviation [27,28]. In this evaluation, a response outlier was determined (3-methyl-1-tetradecylimidazolium chloride) [28].

As can be seen in Tables 5–7, the most adequate values of the correlation coefficient of predicted vs. real values, MPE and

**Table 5**
Regression coefficients and statistics of the fits obtained from MLRs using the verification sample

| | $C$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ | $\alpha_9$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ | $R^2$ | $\sigma$ | MPE (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IPC-81 | 5.767 | 0.023 | −0.093 | 0.252 | −0.030 | −0.070 | 0.351 | 0.222 | 0.392 | 0.219 | $-2.4 \times 10^{-3}$ | 0.584 | 0.286 | 0.867 | 0.07 | 10.6 |
| IPC-81 | – | −0.030 | −0.074 | 0.005 | 0.357 | 0.290 | 0.314 | −0.052 | 0.475 | 0.498 | 0.496 | 0.550 | 1.079 | 0.265 | 0.1 | 22.0 |
| AChE | 2.650 | 0.082 | $3.8 \times 10^{-3}$ | −0.023 | 0.090 | −0.147 | 0.448 | 0.025 | 0.434 | 0.537 | 0.225 | 0.686 | 0.417 | 0.814 | 0.06 | 7.7 |
| AChE | – | 0.058 | 0.017 | −0.137 | 0.268 | 0.018 | 0.430 | −0.100 | 0.473 | 0.664 | 0.454 | 1.589 | 0.781 | 0.452 | 0.1 | 11.3 |

$$\log_{10}\mathrm{EC}_{50} = C + \sum_{i}^{12}\alpha_i \cdot \mathrm{PC}_i.$$

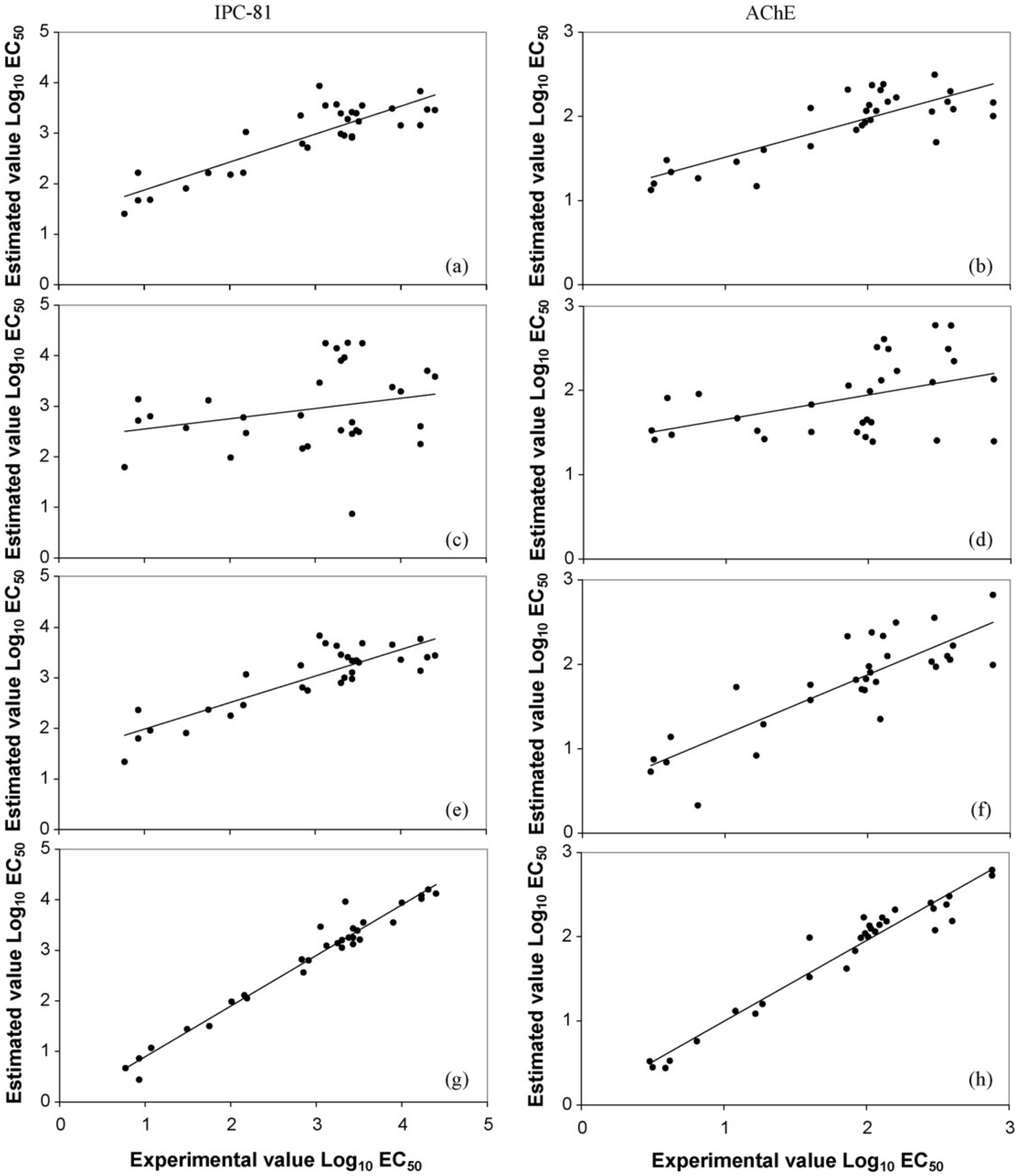

**Fig. 2.** Performance of interpolated models (verification sample); (a and b) MLR with constant; (c and d) MLR without constant; (e and f) RB model; (g and h) MLP model.

**Table 7**
Parameters of the MLP model (with a 95% confidence level)

| Optimized parameters of the NNs | | |
|---|---|---|
| Transfer function | | Sigmoid |
| Training function | | TrainBR |
| Number of input neurons | | 12 |
| Number of hidden neurons | | 2 |
| Number of output neurons | | 2 |
| Lc | | 1 |
| Lcd | | 0.001 |
| Lci | | 2 |
| | IPC-81 | AChE |
| Final prediction error (verification sample) | | |
| MPE (%) | 3.9 | 2.8 |
| $R^2$ | 0.982 | 0.973 |
| $\sigma$ | 0.03 | 0.02 |

standard deviation of both $\log_{10}$ $EC_{50}$ are reached using the MLP model.

Given that a high $R^2$ value does not guarantee that any model fits the data well, in every tested model a residuals analysis of the data was carried out. As can be seen in Fig. 3, there is a correlation between residuals (the difference between the observed value and the corresponding fitted value) from MLR models and both $\log_{10}$ $EC_{50}$, which is in agreement with their high MPE values, Table 5. The least correlation coefficient of the residuals values vs. $\log_{10}$ $EC_{50}$ is reached when the model used is MLP (Fig. 3). From the residuals point of view, the MLP is the most adequate model to estimate the $\log_{10}$ $EC_{50}$ in both IPC-81 and AChE systems in the ILs tested. It is in agreement with the results shown in Sections 3.2 and 3.3.

In every model tested, statistical differences between real and estimated values were calculated using the *p*-value (response of each statistical analysis test), Table 8 [29]. In every model, the simi-

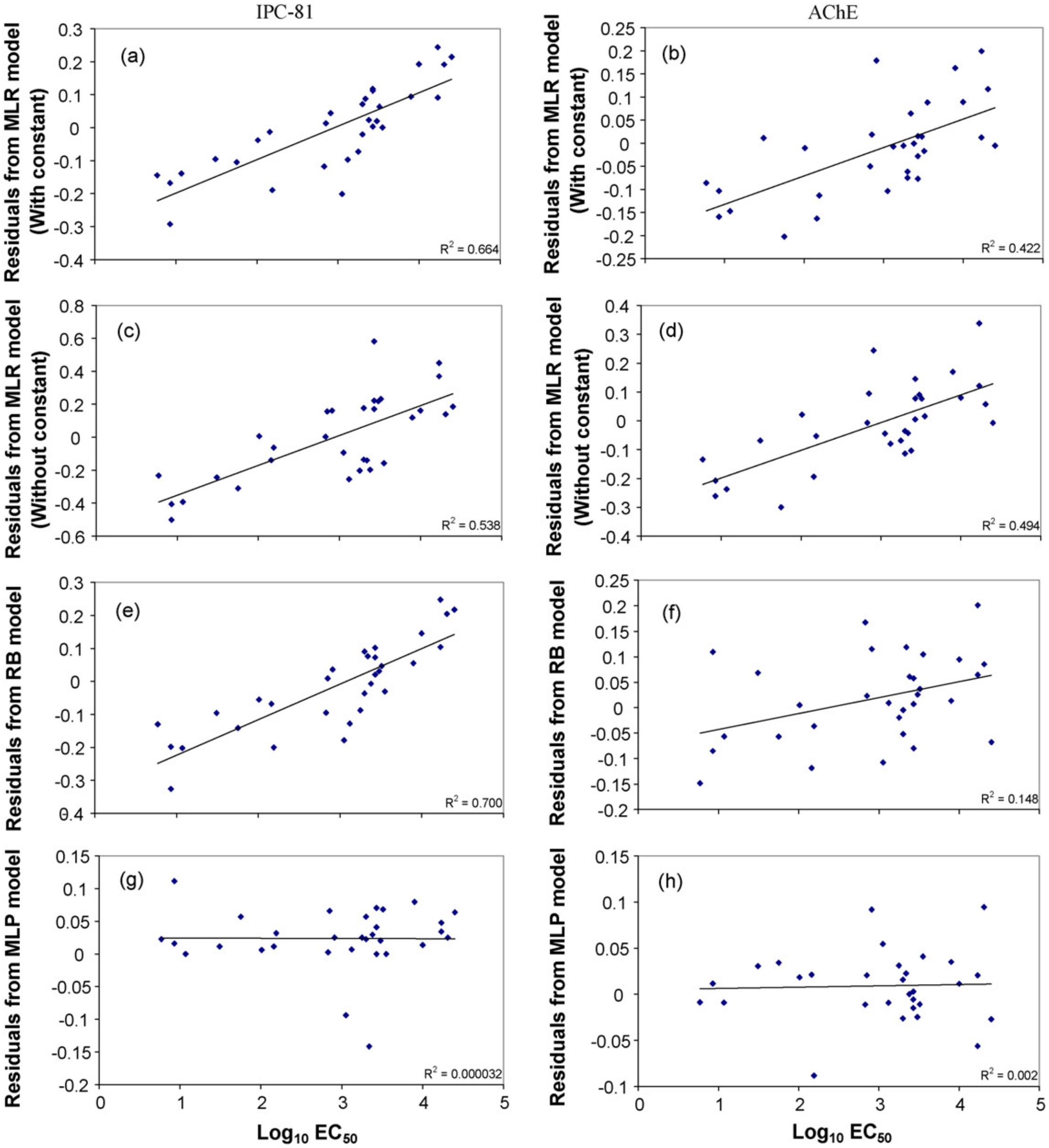

**Fig. 3.** Graphical analysis of residuals from the interpolated models studied (verification sample); (a and b) MLR with constant; (c and d) MLR without constant; (e and f) RB model; (g and h) MLP model.

larity of the real and estimated values was evaluated analyzing their medians by Mann–Withney and Kruskal–Wallis tests. Taking into account their mean *p*-values, the MLP model was the most reliable. The variances of real and estimated databases were also compared by the *F*-test. Due to their *p*-values, the real and estimated values calculated by the MLP model can be considered statistically similar. From the mean of real and estimated values (*t*-test) point of view, the estimations of $\log_{10}$ $EC_{50}$ in both IPC-81 and AChE systems by

**Table 8**
*p*-Values calculated by statistical analysis tests applied to real vs. those estimated values in the verification sample by every applied model (threshold value = 0.05)

| Test type | IPC | | | | AChE | | | |
|---|---|---|---|---|---|---|---|---|
| | Linear models | | Non-linear models | | Linear models | | Non-linear models | |
| | Linear regression | | MLP | RB | Linear regression | | MLP | RB |
| | Without constant | With constant | | | Without constant | With constant | | |
| Mann–Whitney | 0.137 | 0.208 | 0.245 | 0.109 | 0.307 | 0.402 | 0.860 | 0.321 |
| Kolmogorov–Smirnov | 0.147 | 0.079 | 0.607 | 0.079 | 0.079 | 0.019 | 0.607 | 0.413 |
| Kruskal–Wallis | 0.135 | 0.205 | 0.242 | 0.108 | 0.304 | 0.398 | 0.855 | 0.317 |
| *F*-test | 0 | 0.008 | 0.249 | 0 | 0 | 0 | 0.155 | 0.024 |
| *t*-Test | 0.036 | 0.044 | 0.144 | 0.021 | 0.085 | 0.094 | 0.379 | 0.309 |

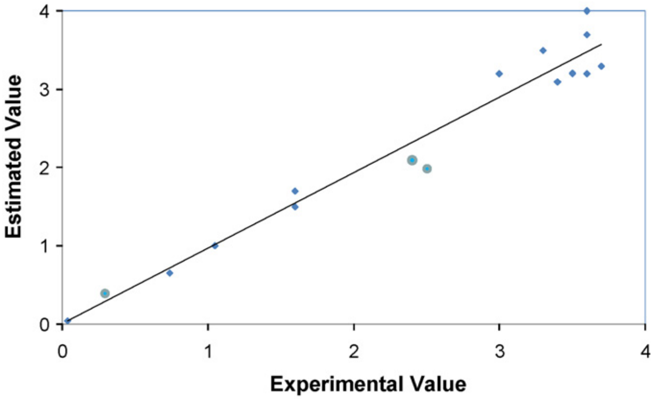

**Fig. 4.** LP model performances using external validation sample (estimations of the $\log_{10}$ $EC_{50}$ in IPC-81 (♦) and AChE (●) systems) [31].

the MLP model can also be considered statistically similar. From the Kolmogorov–Smirnov test point of view, real values and those estimated by non-linear models are statistically similar. Finally, from a central tendency and statistical dispersion analysis point of view, the most adequate model to estimate both $\log_{10}$ $EC_{50}$ is MLP.

### 3.5. *Application of MLP model*

Finally, in order to carry out an external validation of the most adequate model [30], a new validation sample based on experimental data available in literature was employed [31]. Taking into account that the external validation sample range must be within the learning sample range and belong to the same application domain (*vide supra*), the data sets used in the external validation of the MLP model were selected. The mathematical procedure followed was similar to the verification process described above. As can be seen in Fig. 4, the statistical results of estimated vs. real values ($R^2 > 0.94$ and MPE < 9%) are worse than those calculated in the internal validation. These statistical results are to be expected because, as Stasiewicz et al. stated, the structures of ionic liquids tested in both databases are different [31]. Nevertheless, taking into account this chemical difference, this statistical result confirmed that the MLP model is an adequate method to calculate the toxicity values with constitutional descriptors. Since $R^2 > 0.94$, the MLP model presents an acceptable goodness of fit [30]. Taking into account the statistical results of internal and external validation processes, the optimized MLP model has sufficient robustness and predictive capacity to estimate the toxicology values by constitutional descriptors [30].

To sum up, two main types of interpolative models have been tested (linear and non-linear). Taking into account the statistical results shown in this work, depending on the final application of the model, both models could be used. If a faster response is required, a MLR with constant model could be designed and applied using experimental data. Otherwise, if the mathematical complexity of the model is not important or if high accuracy is required, the MLP model should be adequate.

## 4. Conclusion

In this work, four mathematical approaches have been designed to estimate the toxicity of ammonium, imidazolium, morpholinium, phosphonium, piperidinium, pyridinium, pyrrolidinium and quinolinium ionic liquids in *Leukemia Rat Cell Line* and *Acetylcholinesterase* enzyme. The estimations have been carried out using only easily available information (empirical formula and molecular weight of ILs) and no assumptions were taken into account in the design and application of the tested models.

Taken into account the statistical results, the most reliable model was found to be MLP. In the internal validation, its mean MPE and $R^2$ values are less than 3.3% and 0.98, respectively. In the external validation, as the structural chemicals used are different, the statistical results are not as good, and as a result the mean MPE and $R^2$ values are less than 9% and 0.94, respectively. Taking the statistical results into account, this model shows an acceptable goodness of fit, sufficient robustness and an adequate predictive capacity to estimate the toxicity of the ionic liquids by constitutional descriptors.

From a performance error and computational effort point of view, two groups could be made. The simplest model, MLR with constant model, shows worse statistical parameters than those calculated by the MLP model. Although the MLP model requires a more complex calculation process to optimize its parameters, it shows the best statistical results. Definitively, in each case, the final applications of the model will help us to select that which is the most adequate.

## Acknowledgements

The authors are grateful to the Spanish “Ministerio de Ciencia e Innovación” for financial support for project CTQ2006-04644 and a Ramón y Cajal research contract of José S. Torrecilla.

## Appendix A

### A.1. *Principal component analysis description*

Principal component analysis (PCA) is one of the most valuable results from applied linear algebra. PCA is used abundantly in all forms of analysis (from neuroscience to computer graphics), because it is a simple and non-parametric method of extracting relevant information from confusing data sets. PCA provides a technique to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structure that often underlies it. It is based on the assumption that most information about classes is contained in the direction along which the variation is the largest. For a given $p$-dimensional data set $\Phi$, the $m$ principal axes $V_1, V_2, \ldots, V_m$, where $1 \le m \le p$, are orthonormal axes on which the retained variance is maximum in the projected space. Generally, $V_1, V_2, \ldots, V_m$ can be given by the m leading eigenvectors of the sample covariance matrix:

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^{\mathrm{T}}(x_i - \mu) \tag{A.1}$$

where $\mu$ and $N$ are, respectively, the sample mean and the number of samples, and $x_i \in \Phi$.

$$SV_i = gV_i, \quad i \in 1, \ldots, m \tag{A.2}$$

where $g$ is the $i$th largest eigenvalue of $S$. The $m$ principal component of a determined observation vector $x \in \Phi$ is given by

$$y = [y_1, \ldots, y_m] = [V_1^{\mathrm{T}}x, \ldots, V_m^{\mathrm{T}}x] = V^{\mathrm{T}}x \tag{A.3}$$

The $m$ principal components of $x$ are decorrelated in the projected space. In multi-class problems, the variations of data are determined on a global basis, that is, the principal axes are derived from a global covariance matrix:

$$\hat{S} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(x_{ji} - \Pi)(x_{ji} - \Pi)^{\mathrm{T}} \tag{A.4}$$

**Table A1**
Experimental values of $\log_{10} EC_{50}$ in IPC-81 and AChE systems used to design and optimize the models used [2]

| Cation | Anion | IPC-81 | AChE |
|---|---|---|---|
| Ammonium | | | |
| (Ethoxymethyl)ethyldimethylammonium | Chloride | 3.59 | 2.36 |
| (Ethoxymethyl)ethyldimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.8 | 2.3 |
| Tetraethylammonium | Bis[1,2-benzenediolato(2-)-O1,O2]borate | 1.17 | 2.9 |
| Benzyldecyldimethylammonium | Chloride | 0.64 | 0.73 |
| Butylethyldimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.43 | 2.03 |
| Butyltrimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.61 | 2.6 |
| Ethyl(2-ethoxyethyl)dimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.28 | 2.55 |
| Ethyl(2-hydroxyethyl)dimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.7 | 2.59 |
| Ethyl(2-methoxyethyl)dimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.31 | 2.45 |
| Ethyl(3-methoxypropyl)dimethylammonium | Bis(trifluoromethylsulfonyl)amide | 3.54 | 2.92 |
| Tetrabutylammonium | Bromide | 2.25 | 2.3 |
| Imidazolium | | | |
| (Ethoxymethyl)methylimidazolium | Chloride | 3.6 | 2.61 |
| 1-(2-Ethoxyethyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.18 | 2.12 |
| 1-(2-Ethoxyethyl)-3-methylimidazolium | Bromide | 4.14 | 2.27 |
| 1-(2-Hydroxyethyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.76 | 2.88 |
| 1-(2-Methoxyethyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.25 | 2.47 |
| 1-(2-Methoxypropyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.34 | 2.58 |
| 1-(2-Methoxypropyl)-3-methylimidazolium | Chloride | 4.49 | 2.61 |
| 1-(3-Hydroxypropyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.66 | 2.74 |
| 1-(8-Hydroxyoctyl)-3-methylimidazolium | Bromide | 2.36 | 1.28 |
| 1-(Cyanomethyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.9 | 2.88 |
| 1-(Ethoxymethyl)-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 3.2 | 2.45 |
| 1-Butyl-3-ethylimidazolium | Tetrafluoroborate | 3.26 | 2.04 |
| 1-Butyl-3-ethylimidazolium | Trifluoroacetate | 3.31 | 2.01 |
| 1-Butyl-3-ethylimidazolium | Trifluoromethanesulfonate | 3.43 | 2.01 |
| 1-Butyl-3-methylimidazolium | Bis(trifluoromethyl)amide | 2.19 | 1.6 |
| 1-Butyl-3-methylimidazolium | Hexafluoroantimonate | 2.26 | 1.81 |
| 1-Butyl-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 2.68 | 1.96 |
| 1-Butyl-3-methylimidazolium | Trifluoromethanesulfonate | 3.02 | 1.93 |
| 1-Butyl-3-methylimidazolium | Dicyanamide | 3.15 | 1.93 |
| 1-Butyl-3-methylimidazolium | Tetrafluoroborate | 3.12 | 1.98 |
| 1-Butyl-3-methylimidazolium | 2-(2-Methoxyethoxy)ethyl sulfate | 3.16 | 1.99 |
| 1-Butyl-3-methylimidazolium | 1-Methylsulfate | 3.21 | 1.95 |
| 1-Butyl-3-methylimidazolium | 1-Octylsulfate | 3.23 | 1.98 |
| 1-Butyl-3-methylimidazolium | Hexafluorophosphate | 3.1 | 2.15 |
| 1-Butyl-3-methylimidazolium | Hydrogensulfate | 3.29 | 1.97 |
| 1-Butyl-3-methylimidazolium | Toluene-4-sulfonate | 3.29 | 2 |
| 1-Butyl-3-methylimidazolium | Bromide | 3.43 | 1.96 |
| 1-Butyl-3-methylimidazolium | Thiocyanate | 3.42 | 2 |
| 1-Butyl-3-methylimidazolium | Chloride | 3.55 | 1.91 |
| 1-Butyl-3-methylimidazolium | 1-Methanesulfonate | 3.51 | 1.99 |
| 1-Butyl-3-methylimidazolium | Iodide | 3.48 | 2.02 |
| 1-Decyl-3-ethylimidazolium | Bromide | 0.53 | 0.92 |
| 1-Decyl-3-methylimidazolium | Tetrafluoroborate | 0.77 | 1.08 |
| 1-Decyl-3-methylimidazolium | Chloride | 1.34 | 1.09 |
| 1-Decyl-3-methylimidazolium | Hexafluorophosphate | 1.5 | 1.68 |
| 1-Ethyl-3-methylimidazolium | Bis[1,2-benzenediolato(2-)-O1,O2]borate | 1.02 | 2.09 |
| 1-Ethyl-3-methylimidazolium | Bis(pentafluoroethyl)phosphinate | 2.83 | 2.09 |
| 1-Ethyl-3-methylimidazolium | Bis[oxalato(2-)]-borate | 2.93 | 2 |
| 1-Ethyl-3-methylimidazolium | Tetracyanoborate | 3.5 | 1.98 |
| 1-Ethyl-3-methylimidazolium | Tetrafluoroborate | 3.44 | 2.05 |
| 1-Ethyl-3-methylimidazolium | Hexafluorophosphate | 3.92 | 2.05 |
| 1-Ethyl-3-methylimidazolium | 1-Ethylsulfate | 3.93 | 2.07 |
| 1-Ethyl-3-methylimidazolium | Trifluoroacetate | 4 | 2.03 |
| 1-Ethyl-3-methylimidazolium | Toluene-4-sulfonate | 3.81 | 2.22 |
| 1-Ethyl-3-methylimidazolium | Hydrogensulfate | 3.99 | 2.13 |
| 1-Ethyl-3-methylimidazolium | Trifluoromethanesulfonate | 4.09 | 2.13 |
| 1-Ethyl-3-methylimidazolium | Thiocyanate | 4.23 | 2.12 |
| 1-Heptyl-3-methylimidazolium | Hexafluorophosphate | 2.3 | 1.91 |
| 1-Heptyl-3-methylimidazolium | Chloride | 2.53 | 2.07 |
| 1-Heptyl-3-methylimidazolium | Tetrafluoroborate | 2.58 | 2.12 |
| 1-Hexadecyl-3-methylimidazolium | Chloride | -0.19 | 0.68 |
| 1-Hexyl-3-ethylimidazolium | Bromide | 2.01 | 1.77 |
| 1-Hexyl-3-ethylimidazolium | Tetrafluoroborate | 2.26 | 1.84 |
| 1-Hexyl-3-methylimidazolium | 1,2-Benzisothiazolonium 1,1-dioxide | 2.29 | 1.96 |
| 1-Hexyl-3-methylimidazolium | Bis(trifluoromethylsulfonyl)amide | 2.24 | 2.15 |
| 1-Hexyl-3-methylimidazolium | Chloride | 2.85 | 1.92 |
| 1-Hexyl-3-methylimidazolium | Hexafluorophosphate | 2.91 | 1.88 |
| 1-Hexyl-3-methylimidazolium | Tetrafluoroborate | 2.98 | 1.88 |
| 1-Methyl-3-(2-phenylethyl)imidazolium | Hexafluorophosphate | 2.93 | 1.9 |
| 1-Methyl-3-[(4-methylphenyl)methyl]imidazolium | Chloride | 2.64 | 1.86 |
| 1-Methyl-3-[(4-methylphenyl)methyl]imidazolium | Tetrafluoroborate | 2.67 | 2.08 |
| 3-Hexyl-1,2-dimethylimidazolium | Tetrafluoroborate | 1.9 | 1.27 |

Table A1 (*Continued* )

| Cation | Anion | IPC-81 | AChE |
|---|---|---|---|
| 3-Methyl-1-nonylimidazolium | Chloride | 1.4 | 1.36 |
| 3-Methyl-1-nonylimidazolium | Tetrafluoroborate | 1.65 | 1.43 |
| 3-Methyl-1-nonylimidazolium | Hexafluorophosphate | 1.85 | 1.62 |
| 3-Methyl-1-octadecylimidazolium | Chloride | 0.01 | 0.96 |
| 3-Methyl-1-octylimidazolium | Tetrafluoroborate | 1.59 | 1.53 |
| 3-Methyl-1-octylimidazolium | Chloride | 2.01 | 1.6 |
| 3-Methyl-1-octylimidazolium | Bis(trifluoromethylsulfonyl)amide | 1.64 | 2.03 |
| 3-Methyl-1-octylimidazolium | Hexafluorophosphate | 1.96 | 2.03 |
| 3-Methyl-1-propylimidazolium | Tetrafluoroborate | 3.47 | 2.3 |
| 3-Methyl-1-tetradecylimidazolium | Chloride | −0.42 | 0.54 |
| Morpholinium | | | |
| 4-(2-Hydroxyethyl)-4-methylmorpholinium | Bis(trifluoromethylsulfonyl)amide | 3.19 | 2.93 |
| 4-(2-Methoxyethyl)-4-methylmorpholinium | Bis(trifluoromethylsulfonyl)amide | 3.81 | 2.9 |
| 4-(Ethoxymethyl)-4-methylmorpholinium | Bis(trifluoromethylsulfonyl)amide | 3.34 | 2.88 |
| 4-(Ethoxymethyl)-4-methylmorpholinium | Chloride | 3.52 | 2.96 |
| 4-Butyl-4-methylmorpholinium | Bis(trifluoromethylsulfonyl)amide | 3.43 | 2.78 |
| 4-Ethyl-4-methylmorpholinium | Toluene-4-sulfonate | 3.81 | 2.59 |
| Phosphonium | | | |
| Tetrabutylphosphonium | Bromide | 1.66 | 2.61 |
| Tetrabutylphosphonium | Bis[1,2-benzenediolato(2-)-O1,O2]borate | 1.32 | 3.11 |
| Piperidinium | | | |
| 1-(2-Ethoxyethyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.34 | 2.55 |
| 1-(2-Ethoxyethyl)-1-methylpiperidinium | Bromide | 4.31 | 2.6 |
| 1-(2-Hydroxyethyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.65 | 2.34 |
| 1-(2-Hydroxyethyl)-1-methylpiperidinium | Iodide | 4.58 | 2.34 |
| 1-(2-Methoxyethyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.28 | 1.93 |
| 1-(3-Hydroxypropyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.63 | 2.56 |
| 1-(3-Methoxypropyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.27 | 2.27 |
| 1-(3-Methoxypropyl)-1-methylpiperidinium | Chloride | 4.4 | 2.2 |
| 1-(Cyanomethyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 4 | 2.45 |
| 1-(Cyanomethyl)-1-methylpiperidinium | Chloride | 4.58 | 2.43 |
| 1-(Ethoxymethyl)-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.41 | 2.16 |
| 1-Butyl-1-methylpiperidinium | Bis(trifluoromethylsulfonyl)amide | 3.41 | 1.78 |
| 1-Butyl-1-methylpiperidinium | Bromide | 4.03 | 1.83 |
| Pyridinium | | | |
| 1-(2-Ethoxyethyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.26 | 1.48 |
| 1-(2-Ethoxyethyl)pyridinium | Bromide | 4.24 | 1.55 |
| 1-(2-Hydroxyethyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.79 | 2.65 |
| 1-(2-Hydroxyethyl)pyridinium | Iodide | 4.15 | 2.69 |
| 1-(2-Methoxyethyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.19 | 2.09 |
| 1-(3-Hydroxypropyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.55 | 2.56 |
| 1-(3-Methoxypropyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.38 | 2.06 |
| 1-(Ethoxymethyl)pyridinium | Bis(trifluoromethylsulfonyl)amide | 3.12 | 2.14 |
| 1-(Ethoxymethyl)pyridinium | Chloride | 3.32 | 2.06 |
| 1-Butyl-2-methylpyridinium | Tetrafluoroborate | 3.25 | 0.82 |
| 1-Butyl-3,4-dimethylpyridinium | Tetrafluoroborate | 3.02 | 1.1 |
| 1-Butyl-3,5-dimethylpyridinium | Tetrafluoroborate | 3.25 | 1.17 |
| 1-Butyl-3-methylpyridinium | Tetrafluoroborate | 3.3 | 1.27 |
| 1-Butyl-4-methylpyridinium | Tetrafluoroborate | 2.98 | 1.54 |
| 1-Butylpyridinium | Tetrafluoroborate | 3.18 | 1.8 |
| 1-Butylpyridinium | Bromide | 3.9 | 1.77 |
| 1-Cyanomethylpyridinium | Bis(trifluoromethylsulfonyl)amide | 3.5 | 2.51 |
| 1-Cyanomethylpyridinium | Chloride | 3.79 | 2.47 |
| 1-Hexyl-4-methylpyridinium | Tetrafluoroborate | 2.17 | 1.48 |
| 1-Octyl-4-methylpyridinium | Tetrafluoroborate | 1.49 | 1.22 |
| 1-Octyl-4-methylpyridinium | Chloride | 1.63 | 1.11 |
| 1-Octylpyridinium | Chloride | 1.27 | 1.6 |
| 4-(Dimethylamino)-1-butylpyridinium | Bis(trifluoromethylsulfonyl)amide | 1.75 | 0.59 |
| 4-(Dimethylamino)-1-butylpyridinium | Chloride | 1.94 | 0.6 |
| 4-(Dimethylamino)-1-ethylpyridinium | Bis(trifluoromethylsulfonyl)amide | 2.84 | 0.93 |
| 4-(Dimethylamino)-1-ethylpyridinium | Bromide | 2.9 | 0.99 |
| 4-(Dimethylamino)-1-hexylpyridinium | Chloride | 0.93 | 0.5 |
| 4-(Dimethylamino)-1-hexylpyridinium | Bis(trifluoromethylsulfonyl)amide | 0.93 | 0.81 |
| Pyrrolidinium | | | |
| 1-(2-Ethoxyethyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.2 | 2.55 |
| 1-(2-Hydroxyethyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.72 | 2.61 |
| 1-(2-Methoxyethyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.3 | 2.11 |
| 1-(Cyanomethyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.8 | 2.83 |
| 1-(Cyanomethyl)-1-methylpyrrolidinium | Chloride | 4.23 | 2.88 |
| 1-(3-Hydroxypropyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.6 | 2.77 |
| 1-(3-Methoxypropyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.4 | 2.71 |
| 1-(Ethoxymethyl)-1-methylpyrrolidinium | Chloride | 3.05 | 1.86 |
| 1-(Ethoxymethyl)-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.26 | 2.22 |

Table A1 (*Continued* )

| Cation | Anion | IPC-81 | AChE |
|---|---|---|---|
| 1,1-Dihexylpyrrolidinium | Tetrafluoroborate | 1.23 | 2.08 |
| 1-Butyl-1-methylpyrrolidinium | Tetrafluoroborate | 2.9 | 1.91 |
| 1-Butyl-1-methylpyrrolidinium | Bis(trifluoromethylsulfonyl)amide | 3.01 | 2.13 |
| 1-Butyl-1-methylpyrrolidinium | Bromide | 3.77 | 1.93 |
| 1-Butyl-1-methylpyrrolidinium | Dicyanamide | 4.23 | 1.98 |
| 1-Hexyl-1-methylpyrrolidinium | Chloride | 2.91 | 2.48 |
| 1-Methyl-1-octylpyrrolidinium | Tetrafluoroborate | 1.82 | 2.02 |
| 1-Methyl-1-octylpyrrolidinium | Chloride | 2.59 | 2.36 |
| Quinolinium | | | |
| 1-Butylquinolinium | Tetrafluoroborate | 2.16 | 0.62 |
| 1-Butylquinolinium | Bromide | 2.32 | 0.79 |
| 1-Hexylquinolinium | Tetrafluoroborate | 1.07 | 0.48 |
| 1-Octylquinolinium | Tetrafluoroborate | 0.17 | 0.3 |

where $\Pi$, $K$ and $N_j$ are, respectively, the global mean of all samples, number of classes and the number of samples in class $j$. The principal axes are the $m$ leading eigenvectors of $\hat{S}$:

$$\hat{S}V_i = \hat{g}V_i, \quad i \in 1, \ldots, m \tag{A.5}$$

where $\hat{g}$ is the $i$th largest eigenvalue of $\hat{S}$. An assumption made for feature extraction and dimensionality reduction by PCA is that most information of the observation vectors is contained in the subspace spanned by the first $m$ principal axes, where $m < p$. Therefore, each original data vector can be represented by its principal component vector with dimensionality $m$ [32]. All mathematical processes of the PCA technique are summarized in Fig. A1.

In this work, preserving the information of the original data, its dimensionality was reduced following the PCA technique described.

### A.2. Neural networks descriptions

The neural network (NN) is characterised by layered architectures and feed-forward connections between neurons, or back connections. Weights are assigned to these connections between the neurons of one layer and the next. In order to predict with the least possible error, these weights must be optimized.

The calculation process in each neuron of the hidden and output layers consists of transfer and activation functions, Fig. A2. The activation function, Eq. (A.6), means that the input data for each neuron are multiplied by a self-adjustable parameter, $w$, called weight; the result, $x_k$, is fed into a transfer function. The sigmoid function is one of the most commonly used as a transfer function, Eq. (A.7). The calculated value, $y_k$, is the output of the considered neuron, Fig. A2.

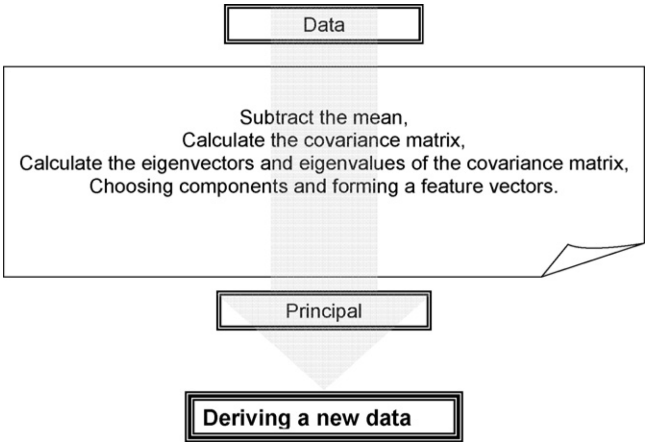
Data
Subtract the mean,
Calculate the covariance matrix,
Calculate the eigenvectors and eigenvalues of the covariance matrix,
Choosing components and forming a feature vectors.
Principal
Deriving a new data

**Fig. A1.** Calculation scheme of the principal component analysis technique.

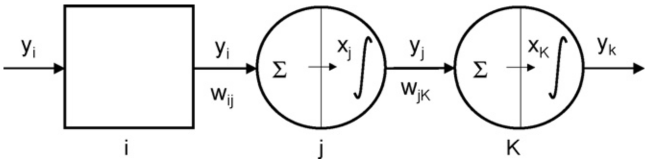

**Fig. A2.** Scheme calculation of MLP model ($y$ = signal, $w$ = weight, $i$-$j$-$k$ = input, hidden and output layer, respectively).

$$x_k = \sum_{j=1} w_{jk} \cdot y_j \tag{A.6}$$

$$y_k = f(x_k) = \left(\frac{1}{1 + e^{-x_k}}\right) \tag{A.7}$$

In Eqs. (A.6) and (A.7), $i$, $j$, and $k$ are the input, hidden and output layers, respectively.

### A.2.1. Learning and verification samples

The learning and verification samples were used to optimize the matrix of weights and to verify the process, respectively. At this stage, the NN uses the input values to predict the output values and no optimization process of weights was carried out. The learning and verification samples were composed of data that characterise the process. Both samples have the same format. These have as many rows as variables necessary to model the process and the same number of columns as the number of vectors to describe the system to be measured.

### A.2.2. MLP model design

The design of the MLP developed involves determining the following factors: training algorithm, NN optimal parameters, topology and transfer function. The MLP used in this work is formed of three layers (input, hidden and output). This topology with a single hidden layer has been considered sufficient to solve similar or more complex problems [13,18]. Moreover, more hidden layers may cause over-fitting [33]. One of the most important stages is the optimization of the matrix of weights. This matrix is optimized using the learning and verification processes.

*Learning process*. The learning sample was presented to the network and a back-propagation algorithm automatically adjusted the weights so that the output response to input vector was as close as possible to the desired response [34]. Each estimation was compared to the corresponding desired value. Then, the estimation error (difference between the estimated and desired values, also called prediction error) was back distributed across the network in a manner that allows the interconnection weights to be modified according to decreases in the estimation error. To optimise the MLP model, several training functions were tested, as described in literature [14]. When the weights were modified, the next data set was fed to the network, and a new estimation was made. The estimation error was calculated again and back distributed across the network for the next modification. Simultaneously, using the verification sample, a verification test was carried out to determine the level of generalization produced by the learning set and to monitor NN over-fitting [35]. When all data of the learning sample were used, an epoch was finished and other one began. To avoid over-fitting of the neural network model, the learning process was repeated while the verification error decreased [36]. A matrix of weights was optimised for each training function [14].

*Verification process*. In this process, the model was tested against the verification sample that had not been included in the neural network learning. The objective of this step was to evaluate the competence of the trained MLP model. For each training function, the matrix of weight optimised above was used. The verification sample was input into the MLP and predicted values were calculated. These were compared with the real values to optimise the parameters of the MLP model by statistical tools. In this process, no corrections of these weights were made and the MLP was only used for prediction.

#### A.2.2.1. *Optimization process of the MLP model.*

In order to optimise the MLP model, two stages were carried out. Firstly, the adequate training function was selected, and then, the main parameters of the MLP model (using the adequate training function) were optimized.

*Selection of the training function*: Using the adequate learning and verification samples, the training function was selected among 14 different functions [14]. To investigate the effect of each training function, all other MLP parameters were set, i.e. the topology was 8 hidden neurons and the others were set as shown in literature [14]. With these conditions and for each training function, a learning process, and then, a verification process were developed. The predicted values were calculated in the verification process. This process was repeated with every training function, and then, all predicted values for each training function were compared individually with the real values. These comparisons were carried out by prediction error, mean square error (MSE), Eq. (A.8), statistical tests and correlation coefficient ($R^2$) (predicted vs. real values). They were carried out to determine if there were significant differences between real data and those predicted by the MLP (at 95% confidence level). To check the null hypothesis (both series are statistically equal), *p*-value (PV) was used. If *p*-value is greater than 0.05, the null hypothesis is fulfilled.

Given that the NN must predict with the highest possible accuracy, the training function selection was carried out to obtain the highest values of PV and $R^2$ of predicted vs. real values and the least MSE, Eq. (A.8).

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N}(r_k - y_k)^2 \tag{A.8}$$

In Eq. (A.8), $N$, $r_k$ and $y_k$ are the number of estimations, and the real and estimated values, respectively.

*Optimization of MLP parameters*: Using the MLP model with the training function selected above, the parameters of the MLP were optimized by a Central Composite Design $2^4$ + star experimental design, where the variables analyzed were the parameters to optimize. As can be seen below, the training function selected in every case was trainBR. The variables analyzed were the hidden neurons number (HNN), which is related to the converging performance of the output error function during the learning process. Too few HNN values would hamper the learning capability of the NN, while too many can cause over-fitting or memorization of the learning sample. The HNN were tested between 1 and 8 neurons [37,38]. The learning coefficient (Lc) controls the degree at which connection weights are modified during the learning phase. The learning coefficient decrease (Lcd) and learning coefficient increase (Lci) control the Lc value. The Lcd and Lc were tested between 0.001 and 1 and Lci between 2 and 100 [18,39]. The responses of experimental design were taken in the learning and verification processes. In the learning process, the epoch number necessary to optimize the matrix of weights and the MSE values were taken. In the verification process, the $R^2$ of predicted vs. real values and mean prediction error were used, Eq. (3).

Learning and verification processes are carried out in each run of the experimental design. Finally, the responses of experimental design were taken. The design was analyzed taking into account that the NN must predict with the least MSE (Eq. (A.8)) and MPE (Eq. (3)), and the $R^2$ must be as close to unity as possible in the lowest iteration number. Given that the learning data are positives, the sigmoid function was selected as transfer function of the MLP model.

### A.2.3. *RB model design*

A RB model was developed to estimate the $\log_{10}\mathrm{EC}_{50}$ of ILs in IPC-81 and AChE systems. In a RB model there are three types of parameters: hidden neurons number, weights and width parameters (spread constant). The hidden neurons number and the matrix of weights are optimized by the RB algorithms [14]. And so, in the RB optimization process, the Spread Constant (SC) is the only parameter to optimize. Therefore, the experimental factor analyzed was SC (between $1 \times 10^{-3}$ and $1 \times 10^{6}$) [14]. The responses were the MPE, standard deviation ($\sigma$) and $R^2$ values. Given that both $\log_{10}\mathrm{EC}_{50}$ must be estimated with the highest possible accuracy, the least error with the highest values of both PV and $R^2$ were the premises to optimize the SC value.

## References

[1] R.F.M. Frade, A. Matias, L.C. Branco, C.A.M. Afonso, C.M.M. Duarte, Effect of ionic liquids on human colon carcinoma HT-29 and CaCo-2 cell lines, Green Chem. 8 (2007) 873–877.

[2] J. Ranke, S. Stolte, R. Stormann, J. Arning, B. Jastorff, Design of sustainable chemical products—the example of ionic liquids, Chem. Rev. (2007) 2183–2206.

[3] S. Stolte, M. Matzke, J. Arning, A. Böschen, W.R. Pitner, U. Welz-Biermann, B. Jastorff, J. Ranke, Effects of different head groups and functionalised side chains on the aquatic toxicity of ionic liquids, Green Chem. 9 (2007) 1170–1179.

[4] F. Stock, J. Hoffmann, J. Ranke, R. Störmann, B. Ondruschka, B. Jastorff, Effects of ionic liquids on the acetylcholinesterase—a structure–activity relationship consideration, Green Chem. 6 (2004) 286–290.

[5] J. Palomar, V.R. Ferro, J.S. Torrecilla, F. Rodríguez, Density and molar volume predictions using COSMO-RS for ionic liquids. An approach to solvents design, Ind. Eng. Chem. Res. 46 (2007) 6041–6048.

[6] P. Luis, I. Ortiz, R. Aldaco, A. Irabien, A novel group contribution method in the development of a QSAR for predicting the toxicity (*Vibrio fischeri* EC50) of ionic liquids, Ecotox. Environ. Saf. 67 (2007) 423–429.

[7] P. Mazzatorta, M. Smiesko, E. Lo Piparo, E. Benfenati, QSAR model for predicting pesticide aquatic toxicity, J. Chem. Inf. Model. 45 (2005) 1767–1774.

[8] C.Y. Zhao, H.X. Zhang, X.Y. Zhang, M.C. Liu, Z.D. Hu, B.T. Fan, Application of support vector machine (SVM) for prediction toxic activity of different data sets, Toxicology 217 (2006) 105–119.

[9] S. Gosav, M. Praisler, D.O. Dorohoi, ANN expert system screening for illicit amphetamines using molecular descriptors, J. Mol. Struct. 834–836 (2007) 188–194.

[10] S. Kara, F. Dirgenali, A system to diagnose atherosclerosis via wavelet transforms, principal component analysis and artificial neural networks, Expert Syst. Appl. 32 (2007) 632–640.

[11] Y.X. Zhang, Artificial neural networks based on principal component analysis input selection for clinical pattern recognition analysis, Talanta 73 (2007) 68–75.

[12] E. Holmes, J.K. Nicholson, G. Tranter, Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks, Chem. Res. Toxicol. 14 (2001) 182–191.

[13] M.C. Palancar, J.M. Aragón, J.S. Torrecilla, pH-Control system based on Artificial Neural Networks, Ind. Eng. Chem. Res. 37 (1998) 2729–2740.

[14] MATLAB User's Guide, v 4.0, Neural Network Toolbox, MathWorks Inc., MA, USA, 2005.

[15] POLHEMUS in statistical analysis using STATGRAPHICS plus, Volume 2, Quality Control and Experimental Design, Statistical Graphics Corporation, Englewood Cliffs, New Jersey, 1999.

[16] K. Heberger, E. Csomos, L.J. Simon-Sarkadi, Principal component and linear discriminant analyses of free amino acids and biogenic amines in Hungarian wines, Agric. Food Chem. 51 (2003) 8055–8060.

[17] A.J. Maren, T. Harston, R.P. Pap, Handbook of Neural Computing Applications, Academic Press Inc., San Diego, 1990, pp. 323–324.

[18] J.S. Torrecilla, A. Fernández, J. García, F. Rodríguez, Determination of 1-ethyl-3-methylimidazolium ethylsulfate ionic liquid and toluene concentration in aqueous solutions by artificial neural network/UV spectroscopy, Ind. Eng. Chem. Res. 46 (2007) 3787–3793.

[19] S. Haykin, Neural, Networks, A Comprehensive Foundation, Prentice Hall Inc., Upper Salad River, New Jersey, 1999.

[20] F. Montañés, T. Fornari, P.J. Martín-Álvarez, A. Montilla, N. Corzo, A. Olano, E. Ibáñez, Selective fractionation of disaccharide mixtures by supercritical $CO_2$ with ethanol as co-solvent, J. Supercrit. Fluids 41 (2007) 61–67.

[21] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.

[22] J.R. Schott, A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix, J. Multivar. Anal. 97 (2006) 827–843.

[23] I.R.C. Hill, M.C. Garnett, F. Bignotti, S.S. Davis, In vitro cytotoxicity of poly(amidoamine)s: relevance to DNA delivery, Biochim. Biophys. Acta 1427 (1999) 161–174.

[24] D.S. Breslow, Biologically active synthetic polymers, Pure Appl. Chem. 46 (1976) 103–113.

[25] J. Ranke, A. Muller, U. Bottin-Weber, F. Stock, S. Stolte, J. Arning, R. Stormann, B. Jastorff, Lipophilicity parameters for ionic liquid cations and their correlation to in vitro cytotoxicity, Ecotox. Environ. Saf. 67 (2007) 430–438.

[26] D.J. Couling, R.J. Bernot, K.M. Docherty, J.K. Dixona, E.J. Maginn, Assessing the factors responsible for ionic liquid toxicity to aquatic organisms via quantitative structure–property relationship modeling, Green Chem. 8 (2006) 82–90.

[27] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for Koc prediction, J. Mol. Graph. 25 (2007) 755–766.

[28] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 670–694.

[29] J.S. Torrecilla, J.M. Aragón, M.C. Palancar, Modeling the drying of a high-moisture solid with an artificial neural network, Ind. Eng. Chem. Res. 44 (2005) 8057–8066.

[30] M. Stasiewicz, E. Mulkiewicz, R. Tomczak-Wandzel, J. Kumirska, E.M. Siedlecka, M. Golebiowski, J. Gajdusb, M. Czerwicka, P. Stepnowski, Assessing toxicity and biodegradation of novel, environmentally benign ionic liquids (1-alkoxymethyl-3-hydroxypyridinium chloride, saccharinate and acesulfamates) on cellular and molecular level, Ecotox. Environ. Saf. 71 (2008) 157–165.

[31] Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models, No. 69, OECD, Series on Testing and Assessment, Organisation of Economic Cooperation and Development, Paris, France, 2007.

[32] X. Wang, K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, Pattern Recognit. 36 (2003) 2429–2439.

[33] R. Ruan, S. Almaer, J. Zhang, Prediction of dough rheological properties using neural networks, Cereal Chem. 72 (1995) 308–311.

[34] H. Ni, S. Gunasekaran, Food quality prediction with neural networks, Food Technol. 52 (1998) 60–65.

[35] A. Ghaffari, H. Abdollahi, M.R. Khoshayand, I. Soltani Bozchalooi, A. Dadgar, M. Rafiee-Tehrani, Performance comparison of neural network training algorithms in modeling of bimodal drug delivery, Int. J. Pharm. 327 (2006) 126–138.

[36] M. Izadifar, F. Abdolahi, Comparison between neural network and mathematical modeling of supercritical $CO_2$ extraction of black pepper essential oil, J. Supercrit. Fluids 38 (2006) 37–43.

[37] M.N. Jadid, D.R. Fairbairn, Neural-network applications in predicting moment–curvature parameters from experimental data, Eng. Appl. Artif. Intell. 9 (1996) 309–319.

[38] Y. Sun, Y. Peng, Y. Chen, A.J. Shukla, Application of artificial neural networks in the design of controlled release drug delivery systems, Adv. Drug Deliv. Rev. 55 (2003) 1201–1215.

[39] V. Vacic, Summary of the Training Functions in Matlab's NN Toolbox, 2005. http://www.cs.ucr.edu/~vladimir/cs171/nn_summary.pdf.