

PCCP

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

Accurate Estimation of Physicochemical Properties of Ternary Mixtures Containing Ionic Liquids via Artificial Neural Networks

John C. Cancilla, Pablo Díaz-Rodríguez, Gemma Matute, José S. Torrecilla*

Departamento de Ingeniería Química, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, 28040-Madrid, Spain.

Abstract

The estimation of density and refractive index of ternary mixtures comprised of the ionic liquid (IL) 1-butyl-3-methylimidazolium tetrafluoroborate, 2-propanol, and water at a fixed temperature of 298.15 K has been attempted through artificial neural networks. The obtained results indicate that the selection of this mathematical approach was a well-suited option. The mean prediction errors obtained, after simulating with a dataset never involved in the training process of the model, were 0.050% and 0.227% for refractive index and density estimation respectively. These accurate results, which have been attained only using the composition of the dissolutions (mass fractions), imply that, most likely, ternary mixtures similar to the one analyzed, can be easily evaluated utilizing this algorithmic tool. In addition, different chemical processes involving ILs can be monitored precisely, and, furthermore, the purity of the compounds in the studied mixtures can be indirectly assessed thanks to the high accuracy of the model.

Key Words: Artificial Neural Network, Ionic Liquid, Density, Refractive Index, Mass Fraction

*Corresponding author. Tel.: +34 91 394 42 44; Fax: +34 91 394 42 43. E-mail address: jstorre@ucm.es.

1. Introduction

In industrial and chemical fields, ecological concerns are becoming a more and more important issue every day. Because of this, the search of environmental-friendly chemical compounds is increasing by the moment. Good examples of these are ionic liquids (ILs), mainly because of their very low vapor pressure (1), which leads to a virtual absence of contaminating air emissions. This chemical property, together with the characteristic thermal stability of ILs (2) and the fact that they are salts which are generally in a liquid state at room temperature (melting points below 100°C), turn these compounds into a safe and low-polluting alternative for many applications. For instance, they are being employed as liquid-liquid extraction media in many separation processes (3,4) or as entrainers to break down azeotropic mixtures (5,6). They are also utilized as catalysts to speed up various organic chemical reactions (7,8) and even as biocatalysts where a room temperature liquid media is usually required (9).

ILs are formed by organic cations, such as imidazolium- or pyridinium-based ones, and inorganic anions, like halides, sulfates, or borates. The combination of different cations and anions, together with the many lateral chains which can be used to functionalize them, enables the creation of millions of custom-made ILs, greatly widening the range of possible applications (10,11).

The fact that ILs are room temperature liquid salts implies that unique properties, which other common salts lack, are contained within them. Additionally, some physicochemical properties, like refractive index or density, highly depend on surrounding conditions (temperature, pressure, and others) (12) and on the purity level of the ILs (13). When impurities such as water or halides are present in ILs, the values of these properties are heavily altered. This circumstance allows defining different physicochemical property values for binary or ternary mixtures involving ILs and other compounds (5,14).

In order to evaluate and create models of the properties of IL mixtures, mathematical methods can be employed such as multiple linear regressions (MLRs). A different possibility is to use more flexible and refreshable artificial neural networks (ANNs), which are algorithms that discover non-linear relations in large databases to empirically estimate results (15,16). These estimations rely on a non-linear interpolation of the results inside the range of the analyzed data, which must have an acceptable

statistical quality and sufficient amount so the designed models end up being accurate, reliable, and applicable in the whole studied range (17).

In this work, models based on MLRs and ANNs have been created to estimate the density and refractive index of a series of ternary mixtures containing different mass fractions of the IL 1-butyl-3-methylimidazolium tetrafluoroborate ([bmim][BF₄]), 2-propanol, and water (Figure 1) at a fixed temperature of 298.15 K (5).

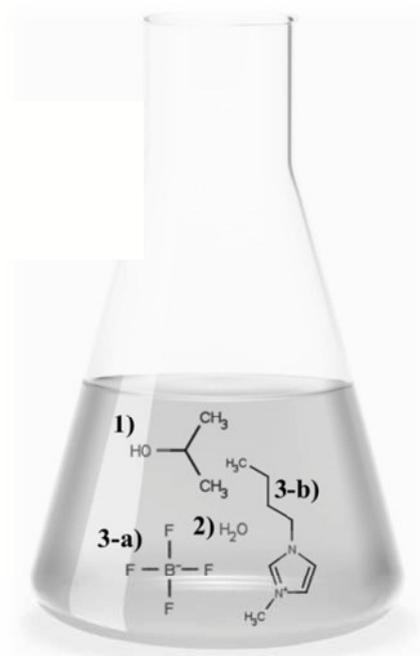


Figure 1. Molecules present in the studied ternary mixture. **1)** 2-propanol; **2)** water; **3-a)** tetrafluoroborate (anion of the IL); **3-b)** 1-butyl-3-methylimidazolium (cation of the IL).

2. Materials and Methods

Information about the database employed and descriptions about the mathematical models utilized, as well as their design, can be seen in the following subsections.

- Database Used for the Mathematical Models

All of the data which was employed to create the ANN-based model, as well as the MLR models, was gathered from the work of Navarro *et al*, 2012 (5). The database

contains information about ternary mixtures of [bmim][BF₄], 2-propanol, and water at a fixed temperature of 298.15 K. All of the compounds used by Navarro *et al.*, 2012 to prepare these mixtures possessed purities greater than 99% (5). The data analyzed and mathematically processed were the mass fractions of the compounds of 72 different dissolutions. These compositions were employed to estimate the density and refractive index values of the ternary mixtures. The database utilized can be found in the **Supplementary Information (SI)** section (**Table S1**).

The 72 data points, which correspond with the dissolutions analyzed, were randomly divided into three groups: training (70% of data), verification (20%), and simulation (10%) datasets. To design the MLR models, the training and verification datasets were combined to develop the models, while the simulation dataset was used to test them. On the other hand, for the ANN model the training set was used to optimize the weights, the verification one to avoid over-fitting effects, and the simulation one to test the final model (*vide infra*). This enables the comparison of both models (MLR versus ANN), as the same data points have been used to test the models. Each data point is labelled in **table S1** with its corresponding dataset.

- Linear Models

Initially, in order to estimate the density and refractive index of the ternary mixtures of [bmim][BF₄], 2-propanol, and water, MLR models have been looked into. This type of modeling is the most common as it is very straightforward and the algorithms behind their design are relatively simple (18). The mass fraction of 2-propanol (**w₁**) and the mass fraction of water (**w₂**) were used as independent variables. These were employed to design two independent MLR models to estimate density (**ρ**; g/cm³) and refractive index (**n_D**) (**Equation 1**). In a further step, the product of both masses (**w₁*w₂**) was also introduced for other MLR models (**Equation 2**). These models will be represented by expressions such as the following (19):

$$y = b + a_1 * w_1 + a_2 * w_2 \quad (1)$$

$$y = b + a_1 * w_1 + a_2 * w_2 + a_3 * w_1 * w_2 \quad (2)$$

Where y represents the dependent variable (n_D or ρ) to be estimated and b , a_1 , a_2 , and a_3 are the parameters of the MLR model.

These mathematical and statistical calculations have been carried out through the software Statgraphics Centurion XVI.

- Artificial Neural Networks

In a second phase, an ANN-based model has been designed with the same goal as the linear models. ANNs are algorithms which rely on determining non-linear relationships present in databases (20). The type of ANN used has been a supervised multilayer perceptron (MLP) (21). It is supervised because it requires target data, in this case, density and refractive index values, to be able to be properly trained and optimized (22).

MLPs are the most regularly employed type of ANN (15). They are manageable mathematical tools which present a series of input nodes as well as calculation centers or neurons in a layer topology. The three layers in a MLP are input, hidden, and output layers. The input layer is formed by nodes, and these are employed to introduce the independent variables which are used to estimate the results (13). The hidden and output layers contain neurons, which are the calculation centers that allow the non-linear interpolation which leads to the estimation of dependent variables, which are designated by the output neurons (23).

Each unit (node or neuron) in every layer of a MLP is connected with all of the units in surrounding layers. These connections are controlled by weighted coefficients or weights, and their optimization is the goal of a MLP. Their role is necessary because not every connection in a MLP has the same relative importance, and, therefore, their proper optimization is vital. In the end, once the weights have been optimized during the training phase, and as long as there are non-linear relations among the data utilized, the accurate estimation of the desired variables should be possible (21).

- Training the Multilayer Perceptron

The equivalent action of training a MLP is optimizing all of the weights which exist within it. To do so, two consecutive mathematical calculations take place in each one of the hidden and output neurons defined. The first one is carried out by an activation function (**Equation 3**), which is in charge of processing the data that enters a neuron.

$$x_k = \sum_{j=1} w_{jk} y_j \quad (3)$$

In the equation above, w_{jk} symbolizes the weight which represents the connection between layers j and k , y_j is the signal that is inputted into a neuron, and x_k is the solution of the activation function (21).

The second calculation is done by a transfer function. Its goal is to limit the range of the resulting values given by a neuron. One of the most common transfer functions is the sigmoid function (**Equation 4**), which provides results between 0 and 1.

$$y_k = \left(\frac{1}{1 + e^{-x_k}} \right) \quad (4)$$

In this function, x_k and y_k are the activation and transfer function solutions respectively (21).

The supervised learning or training process begins when a training dataset is analyzed by the MLP. The weights are then modified so that the values of the outputted results become more similar to the target values. This is attained through **Equations 3** and **4**. After these calculations, the network uses a verification dataset to ensure the generalization capability of the ANN. This verification dataset is not involved in the weight modification and, therefore, the error between the estimated and real values will end up increasing because the weights are being optimized to lower the error from the training dataset instead. Additionally, possible over-fitting effects for a specific database are better avoided when using a verification dataset during the learning process (24).

In addition to this, a third dataset can be used to test the applicability of the optimized ANN. This will be referred to as the simulation dataset, which is not involved in the training process in any matter whatsoever. Once the MLP is optimized, the

simulation dataset is utilized to test it, and its accuracy is then evaluated to compare real and estimated results.

- *Selection and Optimization of Artificial Neural Network Operating Conditions*

In this subsection, together with the **SI** section, the input node, output variable, and training function selection, as well as the hidden neuron number (HNN) and ANN parameter optimization can be found.

· Input Node and Output Neuron Selection

Two input nodes have been defined for the designed MLP: mass fraction of 2-propanol (w_1) and mass fraction of water (w_2). Through the use of these inputs, the ANN is trained to estimate density (ρ ; g/cm³) and refractive index (n_D) values of the mixtures, and, therefore, two output neurons have been defined, one for each of the physicochemical properties studied. The inputted and target data for the supervised ANN can be found in the **SI (Table S1)**.

· Training Function

The training function that has been used for the learning process of the ANN has been the Bayesian regulation function (trainBR). The reason behind this choice is that this training function improves the typical ANN generalization because it updates the weights of the network by analyzing the errors and the sum of the squares of the network weights which allows finding the most important parameters of the ANN and optimizes them. What this implies is that over-fitting ANNs are avoided, and the discovery of the optimum network topology is simplified (24,25).

· Hidden Neuron Number and Artificial Neural Network Parameter Optimization

The HNN optimization is relevant because it ensures that the MLP has reached its full learning potential without over-fitting to the training dataset. More information, as well as the optimization process can be found in the **SI** section (**Table S2**).

On the other hand, a meticulous experimental design based on the “Box-Wilson Central Composite Design 2^3 + star points” has been developed for the following network parameters in order to optimize their values: the Marquardt adjustment parameter (Lc), the decrease factor for Lc (Lcd), and the increase factor for Lc (Lci) (25). The Lc parameter acts as the learning coefficient in the classic back-propagation algorithms (26). Its value is respectively increased or decreased by Lci and Lcd parameters until these changes result in a reduced performance value (25). When this happens, the parameters have acquired their optimal value. A clear explanation about this experimental design, and the results obtained, are shown in the **SI** (**Tables S3** and **S4**). All ANN-related calculations have been done using the software Matlab version 7.0.1.24704 (R14) (25).

3. Results and Discussion

The goal of the MLR and the ANN-based models is to estimate density and refractive index of ternary mixtures of the IL [bmim][BF₄], 2-propanol, and water (**Figure 1**). The results and comparison of both calculation processes are shown in the next subsections.

To evaluate and compare the statistical performances of the models (MLRs and MLP), the R^2 correlation coefficient and the mean prediction error (MPE; **Equation 5**) of the estimated versus real values of the simulation dataset have been calculated (15).

$$MPE = \frac{1}{N} \sum_{k=1}^N \frac{|r_k - y_k|}{r_k} \cdot 100 \quad (5)$$

In this equation, N represents the number of data points in the simulation dataset, r_k is the real already known target value, and y_k stands for the result provided by the linear or non-linear model.

- *Linear Model Development and Performance*

Two different MLR models were designed for each dependent variable estimated. The first ones used two independent variables: the mass fraction of 2-propanol (w_1) and the mass fraction of water (w_2). The second set of models additionally employed the product of both masses ($w_1 * w_2$).

The final resulting MLR models can be seen in **equations 6-9** (**Eq. 6**: two mass fractions to estimate n_D ; **Eq. 7**: two mass fractions to estimate ρ ; **Eq. 8**: two mass fractions and their product to estimate n_D ; **Eq. 9**: two mass fractions and their product to estimate to estimate ρ), while their statistical performance in terms of R^2 and MPE can be found in **table 1**, using the simulation dataset (**Table S1**) to test them.

$$n_D = 1.4181 - 0.0373 * w_1 - 0.0804 * w_2 \quad (6)$$

$$\rho = 1.175 - 0.411 * w_1 - 0.168 * w_2 \quad (7)$$

$$n_D = 1.4197 - 0.0465 * w_1 - 0.0914 * w_2 + 0.0670 * w_1 * w_2 \quad (8)$$

$$\rho = 1.177 - 0.422 * w_1 - 0.181 * w_2 + 0.081 * w_1 * w_2 \quad (9)$$

Table 1. Statistical performance of the MLR models in terms of MPE and R^2 for the estimation of the studied properties. Results obtained after testing the models with the simulation dataset (**Table S1**).

Variables	Dependent Variable	MPE (%)	R^2
w_1 & w_2	¹ n_D	0.085	0.993
	² ρ	0.659	0.993
w_1 , w_2 , & $w_1 * w_2$	³ n_D	0.049	0.999
	⁴ ρ	0.685	0.994

¹Equation 6

²Equation 7

³Equation 8

⁴Equation 9

These statistical results reveal that MLRs are suited to fulfil these estimations as accurate results have been obtained. Furthermore, the use of the product of both variables allows lowering the MPE for the estimation of refractive index to nearly half, while maintaining stable the performance for density.

- *Artificial Neural Network Topology and Parameter Optimization*

In this section, an ANN-based model has been developed to perform the same task as the MLR models. It will be used to determine whether this non-linear approach is better fit for this purpose. The different optimization steps that are required to design a reliable MLP model are shown.

· Topology of the Neural Network

The designed MLP has two independent variables or input nodes that represent the mass fractions of 2-propanol (w_1) and water (w_2) of the ternary mixtures (*vide supra*). The selected HNN was eight because it was the tested case that offered best results in terms of verification MPEs (SI). Finally, the output layer has two neurons which are used to estimate density (ρ ; g/cm^3) and refractive index (n_D) of the ternary mixtures (Figure 2).

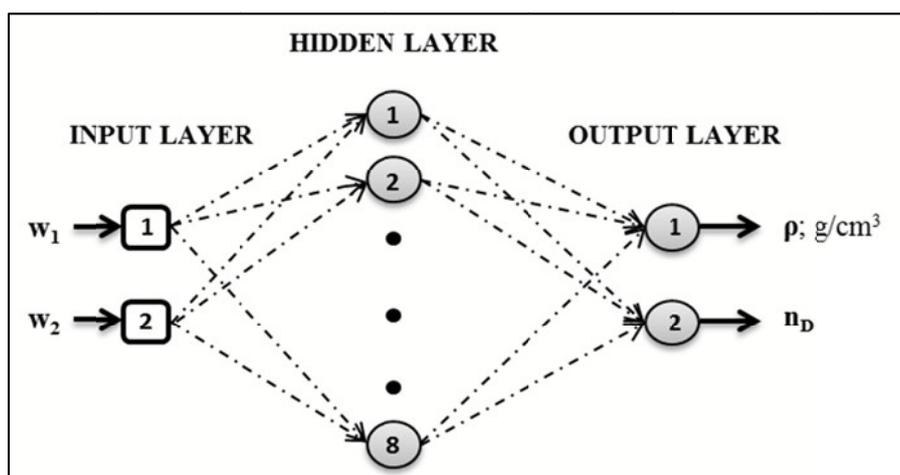


Figure 2. Topology of the designed MLP. Two input nodes (w_1 and w_2), eight hidden neurons, and two output neurons (ρ and n_D) can be seen.

· Artificial Neural Network Parameter Selection and Optimization

After defining the topology, different ANN parameters were selected (transfer and training function) and others optimized (L_c , L_{cd} , and L_{ci}). The results can be seen in **Table 2**. The entire process followed before this selection and optimization process can be found in the SI section (**Tables S2, S3, and S4**).

Table 2. Selected and optimized neural network parameters for the designed MLP.

	Parameter	Value
Selection	Transfer function	Sigmoid

	Training function	TrainBR
Optimization	Lc	0.40
	Lcd	1.00
	Lci	100

- Estimation of Density and Refractive Index through Artificial Neural Networks

Once the desired training and transfer functions have been defined, the optimization of the necessary ANN parameters calculated, and the topology established, the MLP-based model is ready to offer an estimation of the two physicochemical properties studied (ρ and n_D) of the ternary mixtures formed by the IL [bmim][BF₄], 2-propanol, and water (**Figure 1**).

The estimations obtained are result of testing the designed MLP with the simulation dataset (**Table S1**), which has never been involved in the training or verification process of the ANN. Therefore, employing this dataset to validate the model allows the evaluation of the generalization capability and applicability of the mathematical tool. The statistical performance of the MLP can be seen in **table 3**, where the MPE for the n_D (n_D MPE) and for the ρ (ρ MPE), as well as their corresponding R^2 values ($n_D R^2$ and ρR^2 , respectively), are shown.

Table 3. Statistical performance of the MLP in terms of MPE and R^2 for the estimation of the studied properties. Results obtained after testing the ANN model with the simulation dataset (**Table S1**).

n_D MPE (%)	$n_D R^2$	ρ MPE (%)	ρR^2
0.050	0.998	0.227	0.999

As can be seen, the MPEs obtained (0.050% for n_D and 0.227% for ρ) and the high R^2 values (0.998 for n_D and 0.999 for ρ) indicate that, in this case, a very precise estimation of both physicochemical properties is achievable by non-linear interpolation with ANNs.

- Linear Versus Non-Linear Model Comparison

In this subsection, the statistical performance of the MLR models and of the MLP will be compared. As can be found in **tables 1** and **3**, the results provided by the linear and non-linear models are shown respectively. Firstly, when comparing the MLR

that uses the same amount of variables as the MLP (w_1 and w_2), it can be seen that the estimation MPEs provided by the MLR almost double for refractive index (0.085% vs. 0.050%) and nearly triple for density (0.659% vs. 0.227%), giving a clear advantage to the non-linear model. On the other hand, the comparison of the MLR model that uses w_1 , w_2 , and w_1*w_2 with the MLP shows that the estimation of the refractive index is practically as accurate for both as MPEs are very similar (0.049% vs. 0.050%), while the error for density is essentially tripled by the linear alternative (0.685% vs. 0.227%). In both comparisons, the non-linear MLP seems to be a better modeling selection as MPEs are lower. In addition, the ANN model offers the ability to refresh and update its weights with new samples which strict linear models do not. This flexibility that MLPs possess implies a great advantage that MLRs lack. Nonetheless, MLRs compensate this deficiency by being very simple and straightforward to design, requiring lower dataset preparation time and computational power.

To sum up, it is possible to accurately estimate density and refractive index values of ternary mixtures of [bmim][BF₄], 2-propanol, and water, at a fixed temperature, using MLRs and ANNs by only knowing the mass fractions of the analyzed dissolutions. For the case of ANNs, as they are updatable and flexible algorithms, it is more than likely that these results can be extrapolated to many other mixtures, virtually resulting in an easy and widely applicable tool to study and precisely estimate physicochemical properties thus allowing the evaluation of determined chemical operations. Additionally, the estimation errors obtained by the MLP are sufficiently low to permit an indirect assessment of the purity level of the compounds in the mixture.

4. Conclusion

A flexible and updatable mathematical tool based on ANNs has been designed to estimate density and refractive index values of ternary mixtures formed by the IL [bmim][BF₄], 2-propanol, and water at a fixed temperature of 298.15 K. The only required information of the dissolutions was the composition (mass fraction) of its components. This tool was compared to a set of MLR models, and it was shown that the non-linear approach was better suited for this estimation. The results obtained from the MLP model created, after testing it with a dataset that was never employed to design the

network (not used in training or verification phases), offered extremely accurate results in terms of property estimation (n_D MPE = 0.050% and ρ MPE = 0.227%), which, in the end, allows the evaluation of various chemical processes and an indirect control of the purity of the involved compounds. Moreover, this mathematical approach could probably be employed for numerous other binary and ternary mixtures, as long as sufficient data with a high enough quality is available.

Acknowledgements

The research leading to these results has achieved funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. HEALTH-F4-2011-258868.

References

- (1) N.V. Plechkova, K.R. Seddon, *Chemical Society Reviews*, 2008, **37**, 123.
- (2) V. Kamavaram, R.G. Reddy, *Journal of Thermal Sciences*, 2008, **47**, 773.
- (3) J.G. Huddleston, H.D. Willauer, R.P. Swatloski, A.E. Visser, R.D. Rogers, *Chemical Communications*, 1998, **16**, 1765.
- (4) A.E. Visser, R.P. Swatloski, W.M. Reichert, R. Mayton, S. Sheff, A. Wierzbicki, J.H. Davis, R.D. Rogers, *Chemical Communications*, 2001, **1**, 135.
- (5) P. Navarro, M. Larriba, S. García, J. García, F. Rodríguez, *Journal of Chemical Engineering Data*, 2012, **57**, 1165.
- (6) L. Zhang, J. Han, D. Deng, J. Ji, *Fluid Phase Equilibria*, 2007, **255**, 179.
- (7) K.P. Boroujeni, P. Ghasemi, *Catalysis Communications* 2013, **37**, 50.
- (8) D. Li, H. Zang, C. Wu, N. Yu, *Ultrasonics Sonochemistry*, 2013, **20**, 1144.
- (9) F. Van Rantwijk, R.A. Sheldon, *Chemical Reviews*, 2007 **107**, 2757.
- (10) J. Palomar, J.S. Torrecilla, V.R. Ferro, F. Rodríguez, *Industrial & Engineering Chemistry Research*, 2008, **47**, 4523.
- (11) J. Palomar, J.S. Torrecilla, V.R. Ferro, F. Rodríguez, *Industrial & Engineering Chemistry Research*, 2009, **48**, 2257.
- (12) H. Rodríguez, J.F. Brennecke, *Journal of Chemical Engineering Data*, 2006, **51**, 2145.
- (13) J.S. Torrecilla, C. Tortuero, J.C. Cancilla, P. Díaz-Rodríguez, *Talanta*, 2013, **113**, 93.
- (14) P. Navia, J. Troncoso, L. Romani, *Journal of Chemical Engineering Data*, 2007, **52**, 1369.
- (15) J.C. Cancilla, J.S. Torrecilla, G. Matute, *Current Biochemical Engineering*, 2014, **1**, 25.
- (16) E.B. Gueguim-Kana, J.K. Oloke, A. Lateef, M.O. Adesiyun, *Renew. Energ.*, 2012, **46**, 276.
- (17) J.S. Torrecilla, M. Deetlefs, K.R. Seddon, F. Rodríguez, *Physical Chemistry Chemical Physics*, 2008, **10**, 5114.
- (18) P. Díaz-Rodríguez, J.C. Cancilla, N.V. Plechkova, G. Matute, K.R. Seddon, J.S. Torrecilla, *Physical Chemistry Chemical Physics*, 2014, **16**, 128.
- (19) J.S. Torrecilla, J. Palomar, J. García, E. Rojo, F. Rodríguez, *Chemometrics and Intelligent Laboratory Systems*, 2008, **93**, 149.
- (20) J.S. Torrecilla, J.C. Cancilla, G. Matute, P. Díaz-Rodríguez, *International Journal of Food Science & Technology*, 2013, **48**, 2528.
- (21) K. Knoerzer, P. Juliano, P. Roupas, C. Versteeg, *Innovative Food Processing Technologies: Advances in Multiphysics Simulation*. Wiley-Blackwell; Oxford 2011.

- (22) D.R. Hush, B.G. Horne, *Signal Processing Magazine, IEEE*, 1993, **10**, 38.
- (23) J.S. Torrecilla, C. Tortuero, J.C. Cancilla, P. Díaz-Rodríguez, *Talanta*, 2013, **116**, 122.
- (24) J.S. Torrecilla, J.M. Aragón, M.C. Palancar, *Industrial & Engineering Chemistry Research*, 2008, **47**, 7072.
- (25) H. Demuth, M. Beale, M. Hagan, *Neural Network Toolbox for Use with MATLAB® User's Guide*. Version 4.0.6. Ninth printing Revised for Version 4.0.6 (Release 14SP3), Natick, MA **2005**.
- (26) M.C. Palancar, J.M. Aragon, J.S. Torrecilla, *Industrial & Engineering Chemistry Research*, 1998, **37**, 2729.