RESEARCH ARTICLE

Prediction of Phase Composition Profile of Three-Compound Mixtures in Liquid-Liquid Equilibrium: A Chemoinformatics Approach

Gonçalo V. S. M. Carrera*[a], Mariana L. Cruz[a], Kyrylo Klimenko[a], José M. S. S. Esperança[a], João Aires-de-Sousa[a]

[a] Dr. G. V. S. M. Carrera*, Dr. M, L. Cruz,
Dr. K, Klimenko, Dr. J. M. S. S. Esperança,
Prof. Dr. João Aires-de-Sousa
Chemistry Department LAQV-REQUIMTE,
NOVA School of Science and Technology
2829-516 Caparica, Portugal
E-mail: goncalo.carrera@fct.unl.pt
Homepage URL: https://laqv.requimte.pt/people/89-goncalo_valente_da_silva_marino_carrera

Supporting information for this article is given via a link at the end of the document.

Abstract: Machine learning models were developed to predict the composition profile of a three-compound mixture in Liquid-Liquid Equilibrium (LLE), given the global composition at certain temperature and pressure. A chemoinformatics approach was explored, based on the MOLMAP technology to encode molecules and mixtures. The chemical systems involved an ionic liquid (IL) and two organic molecules. Two complementary models have been optimized for the IL-rich and IL-poor phases. The two global optimized models are highly accurate, and were validated with independent test sets, where combinations of molecule1 + molecule2 + IL are different from those in the training set. These results highlight the MOLMAP encoding scheme, based on atomic properties to train models that learn relationships between features of complex multi-component chemical systems and their profile of phase compositions.

Introduction

The present global environmental, social and economic context requires sustainable and efficient industrial chemical processes. The knowledge of the phase equilibrium profile of a chemical system is of outmost importance for the design of those processes. The virtual infinite combinations of chemicals, molar fractions, temperature and pressure make optimisation by trial and error an impossible task. Machine-learning approaches progressively acquired relevance with the increasing amount of available experimental data. [1] However, database curation is essential in order to construct robust models [2] revealing relationships between the characteristics of a chemical system and the property/activity of interest. The unitary operations, extractions, extractive distillations and distillations require a quantitative knowledge of phase behavior and machine learning estimations can assist in new situations. Ionic liquids (ILs) have been tested in order to promote those unitary operations. Characteristics such as neglectable vapour pressure, thermal stability and possibility of tuning physico-chemical properties with the judicious choice of cation and anion, make this class of chemicals promising alternatives to conventional volatile organic solvents. Predictive methods for phase behaviour have been tested with different degrees of generalization, namely for specific azeotrope-breaking systems, [3-7] separation of alkanes from aromatics, [8] separation of alcohols from alkanes and aqueous broths, [9] desulfurization of fuels, [10] constitution of aqueous biphasic systems, [11] mixtures of ILs and amines on CO_2 capture. [12] Some of those studies comprise ternary chemical systems with a restricted applicability. [13-17]

The present study aims to answer a straightforward question: what is the composition profile of a generical mixture of three component chemicals in LLE, at a certain global composition and temperature/pressure? The three-compound mixture includes an IL and two molecules. This is a chemoinformatic approach based on molecular maps of atom-level properties (MOLMAPs) codification system. [18] MOLMAPs are derived from the pattern of a *Kohonen* neural network activation by atoms, bonds or atomic inter-component interactions of a chemical system, based on atomic properties. This approach enables to compare systems of different nature, including different number of components, and can take into account the molar fractions of each chemical. MOLMAPs have been applied in different studies such as the classification of chemical reactions without assignment of reaction centers, [18] conversion of descriptor's dimensionality in QSAR applications, [19] chemical reactivity evaluation from databases with no negative data, [20] classification of metabolic reactions, [21, 22] mutagenicity prediction, [23] QSAR analysis of phenolic antioxidants, [24] viscosity classification [25, 26] and gas solubility in ILs. [27]

Here the machine learning protocol involved Random Forest models [28,29] that receive information about the mixture (molecules and their relative amount), the temperature and the pressure, and predict the molar fraction of one component of the mixture (one of the two molecules) in a specific phase. Different models were trained for the IL-rich and IL-poor phases. In order to specify the molecule to be predicted, its MOLMAP code was concatenated to the MOLMAP code of the whole system.

Therefore, the independent variables are the MOLMAP of the whole system (with components weighted according to their molar fractions), the MOLMAP of one of the molecules, the temperature and pressure; the dependent variable is the molar fraction of the latter molecule. The models were trained with data sets including different ILs and multiple classes of molecules retrieved from the ILThermo database. [30]

Results and Discussion

Random Forest models were explored in order to predict the composition profile of each phase in LLE, given a mixture of one ionic liquid and two organic molecules with known global composition, at a certain temperature and pressure.

A given chemical system, a mixture of an ionic liquid and two organic molecules, at certain conditions of temperature and pressure, is associated to multiple tie-lines. Each tie-line links an IL-poor and IL-rich phase's composition profiles in equilibrium. A specific tie-line is characterized in this work by five randomly generated global composition points. A given datapoint comprises a given global composition point of a specific tie-line for a three-compound system.

Several schemes were explored for the codification of the molecules with the MOLMAP descriptors, namely the size of the MOLMAP for the mixture and for the organic molecule to be predicted, as well as the contributions of neurons around the winning neuron to the activation pattern. The optimized models (IL-poor and IL-rich phases) use as encoding parameters:

1- MOLMAPs for mixtures with size 25x25=625 descriptors with a pattern of activation weighted by 1-0.75-0.5-0.25 for the winning neuron and for the neurons up to a distance of 3.
2- MOLMAPs for the organic molecule with size 25x25=625 descriptors with pattern of activation of 1-0.75-0.5-0.25.

3- Temperature and pressure.

The built models are based on the Random Forest (RF) algorithm considering its ability to handle large numbers of descriptors and its typical good accuracy. The value of *mtry* has been set to 200 in order to assure the contribution of relevant descriptors, such as temperature, to many trees of the ensemble. The high number of predictors, 1000 trees, with a certain degree of variability, increases the probability for an accurate average prediction. [31] All these aspects of the RF algorithm are fundamental to process large data sets in a straightforward way.

The global results obtained for the optimized IL-rich and IL-poor phase-models are highlighted in Table 1. A high accuracy was observed in the internal validation with the training set (estimation by the out-of-bag RF procedure) as well as in the validation with the optimization and test sets. The MAE errors in Table 1 are below 5.5% when compared with the range of molar fractions (0-1) All combinations of four components in the optimization and test sets are different from any combination in the training set. This simulates predictions for unknown mixtures. The models predict the molar fraction of one of the organic molecules in the mixture. Therefore, predictions were obtained for the two molecules by submitting each of them separately (results in Table 1 for "2mol"). The molar fraction of the IL could be calculated from the molar fractions of the two organic molecules - the results including this prediction are in Table 1 under "2mol+IL"). The arithmetic operation to estimate the IL composition is followed by a zero-value assignment for a given datapoint with negative molar fraction. Finally, the sum of the values of molar fraction of the two organic molecules and the ionic liquid are normalized to one.

Table 1. Global results the RF predictions of molar fractions in LLE with the optimized models. Predictions were obtained for the two organic molecules in the mixtures (2mol) and also including the molar fraction of the IL calculated from the predicted values for the other two compounds (2mol+IL).

	IL-Poor Phase			IL-Rich Phase	
	Training Set (Out of Bag)				
	2mol	**2mol+IL**		**2mol**	**2mol+IL**
R^2	0.9822	0.9873	**R^2**	0.9817	0.9875
MAE	0.0229	0.0189	**MAE**	0.0153	0.0178
RMS	0.0480	0.0430	**RMS**	0.0323	0.0347
	Optimization Set				
	2mol	**2mol+IL**		**2mol**	**2mol+IL**
R^2	0.9540	0.9706	**R^2**	0.9067	0.9500
MAE	0.0438	0.0346	**MAE**	0.0345	0.0411
RMS	0.0740	0.0642	**RMS**	0.0635	0.0718
	Test Set				
	2mol	**2mol+IL**		**2mol**	**2mol+IL**
R^2	0.9358	0.9571	**R^2**	0.9043	0.9398
MAE	0.0516	0.0426	**MAE**	0.0406	0.0483
RMS	0.0876	0.0782	**RMS**	0.0669	0.0770

R^2 – square of the Pearson Coefficient. MAE - Mean Absolute Error. RMS - Root Mean Square Error.

Our optimized models have been validated by randomization of mole fractions of the training set for IL-poor and IL-rich phases. Those randomized models have been applied to the independent optimization and test sets. The results for the IL-poor randomized model revealed a MAE of 0.2991/0.2136 and RMS 0.3422/0.2817 for the optimization set (2mol/2mol+IL). Similarly, the test set yielded a MAE of 0.3012/0.2129 and an RMS of 0.3425/0.2809. For the IL-rich phase model the MAE was 0.1943/0.2248 and the RMS 0.2238/0.2624 for the optimization set, while a MAE of 0.1921/0.2141 and RMS of 0.2248/0.2515 was obtained for the test set. The zero-model validation, based on the average of the molar fractions in the training set, has been carried out. The results for the IL-poor phase are MAE 0.2956/0.2001 and RMS 0.3352/0.2742 (optimization set) and MAE of 0.2968/0.2002 and RMS of 0.3357/0.2742 (test set). The accuracy obtained for the IL-rich phase model (optimization set) was MAE of 0.1797/0.2059 and RMS of 0.2099/0.2399; for test set the MAE was 0.1806/0.2002 and the RMS 0.2168/0.2368. Both validation tests demonstrate the significance of the obtained optimized models.The Random Forest seed random control parameter has been tested in order to assess the impact of random fluctuations in the predictions. Three alternative values of *seed* yielded results with deviations in R^2, MAE and RMS below 1.5%. The results of all these forms of validation in combination with Table 1 show that that there's a significant relationship between the codification of a given chemical system's datapoint and the property of interest, the phase behaviour profile.

The use of MOLMAP encoding technology assures a generical approach to encode mixtures of organic molecules based on the properties of atoms. It can also distinguish similar atoms in different structural environments and detect similarities between different atoms with similar profiles of atomic properties. This framework enables to encode and compare chemical systems with different number of components / different nature using a fixed size matrix/vector, accounting for the proportions of components in the mixture. The results here obtained with the optimized MOLMAPs, Table 1, were compared with models trained with reference fingerprints obtained with the Padel software. [32] The encoding strategy was similar to that implemented with MOLMAPs, yielding attributes that represent a)

the generical three compound mixture (two molecules + IL) by 307 descriptors, b) the molecule to be predicted (307 descriptors), c) temperature and d) pressure. Each fingerprint (a count descriptor) is weighted by the respective component global molar fraction (organic molecule 1 + organic molecule 2 + cation + anion). The sum of all the mixture global component molar fractions is normalized to the value of two. The weighted fingerprints of the components are summed to obtain the fingerprint of the mixture. The organic molecule is also explicitly and independently represented by its molar fraction-normalized fingerprint. Each datapoint is represented by 614 structural descriptors + Temperature + Pressure.

The results are presented in Table 2. A slightly worse performance was observed than with MOLMAPs, especially regarding the test set.

Differently from MOLMAPs, the fingerprint count descriptors are based on fixed fragments. MOLMAPs offer the possibility to account for more general features during the training of a model and to obtain predictions for chemical systems with different structural configuration, however with related profiles of atomic properties.

Table 2. Results obtained with RF models trained with reference fingerprint count-based descriptors to predict molar fractions of components in LLE.

	IL-Poor Phase			IL-Rich Phase	
			Training Set (Out of Bag)		
	2mol	2mol+IL		2mol	2mol+IL
R^2	0.9749	0.9826	R^2	0.9761	0.9830
MAE	0.0282	0.0232	MAE	0.0188	0.0225
RMS	0.0563	0.0500	RMS	0.0368	0.0416
	2mol	2mol+IL		2mol	2mol+IL
			Optimization Set		
R^2	0.9557	0.9713	R^2	0.8978	0.9449
MAE	0.0424	0.0327	MAE	0.0370	0.0446
RMS	0.0719	0.0627	RMS	0.0666	0.0754
	2mol	2mol+IL		2mol	2mol+IL
			Test Set		
R^2	0.8891	0.9303	R^2	0.8840	0.9246
MAE	0.0572	0.0448	MAE	0.0435	0.0523
RMS	0.1121	0.0961	RMS	0.0740	0.0859

The use of k-nearest neighbour algorithm, knn, comprising a single neighbour, [33] was tested in order to have an indication if overfiting/data memorization is ocurring in the optimized MOLMAP/RF approach. It has been used the scikit-learn platform to attain knn regressor with k=1. The remaining parameters have been selected as default. [34] The results obtained with this regressor (Table 3) are significantly worse than the optimized MOLMAP/RF models (Table 1), which is an indication of generalization ability of our approach.

Table 3. Results obtained with k-nn (k=1) algorithm and the optimized MOLMAP descritors in order to predict molar fractions of components in LLE.

	IL-Poor-Phase			IL-Rich-Phase	
			Optimization Set		
	2mol	2mol+IL		2mol	2mol+IL
R^2	0.8812	0.8927	R^2	0.8189	0.8996
MAE	0.0605	0.0502	MAE	0.051	0.0604
RMS	0.1192	0.1188	RMS	0.0915	0.1038
			Test Set		
	2mol	2mol+IL		2mol	2mol+IL
R^2	0.8482	0.866	R^2	0.7999	0.8752
MAE	0.0661	0.0538	MAE	0.0571	0.068
RMS	0.1361	0.1336	RMS	0.1001	0.1134

The previous works on ternary systems present restricted applicability. [13-17] One reference study is a review on different COSMO-RS models applied to LLE of aromatic/aliphatic partition in ionic liquids. [13] The models show correct trends on structural variability of aromatic, aliphatic, cation and anion, but low quantitative accuracy. The models here shown present excellent quantitative accuracy for alkane/aromatic mixtures in ionic liquids - Tables 4 and 5. One of the reasons for the excellent results is the large size of this class of mixtures in the training set (Figure 1). Another study concerned the LLE of thiophene/alkane mixtures in ionic liquids. [16] Artificial neural networks and support vector machine-based algorithms have been tested. The results are good with limited applicability. Our approach is centred on global models for different classes of organic molecule-based mixtures, and the performance obtained for alkane/miscellaneous is excellent, Tables 4 and 5, (thiophene is included in the miscellaneous category). Alkane/miscellaneous is the third most represented class of organic molecule-based mixtures (Figure 1).

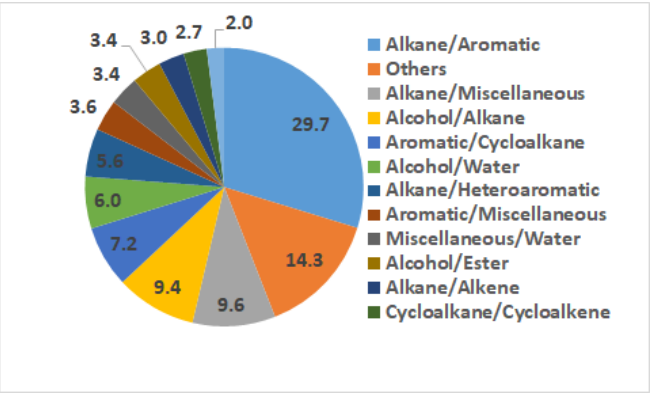


Figure 1. Percentage of chemical systems organized by molecule types in the training set.

Simoni et al. [17] used an asymmetric non-random two-liquid model (NRTL)/eNRTL for mixtures of solvent, water and ionic liquid in LLE (molecular and electrolyte-based phases), with diverse alcohols as solvents. The approach offers better performance than symmetric evaluation. Our models present excellent accuracy results for mixtures of alcohol/water, Tables 4

and 5, even considering that this class of mixtures is moderately represented in the training set, Figure 1.

Table 4. Distribution of class types in the mixtures of molecules and results of RF predictions for each type (optimization set).

		IL-poor phase	IL-rich phase
Mixture Type	**%**	**MAE**	**MAE**
Alkane/Aromatic	43.36	0.0305	0.027
Alcohol/Alkane	6.95	0.064	0.0341
Alkane/Miscellaneous	6.5	0.0413	0.0342
Alcohol/Water	6.41	0.073	0.0413
Aromatic/Cycloalkane	6.23	0.0285	0.0194
Alkane/Heteroaromatic	5.61	0.0409	0.0283
Alkane/Alkene	4.27	0.0573	0.0215
Miscellaneous/Water	3.65	0.0192	0.0795
Cycloalkane/Cycloalkene	3.65	0.0459	0.0125
Alcohol/Ester	2.85	0.0751	0.1034
Aromatic/Miscellaneous	1.96	0.0736	0.0648
Alkane/Ester	1.87	0.0487	0.0504
Alkane/Cycloalkane	1.78	0.0797	0.0151
Cycloalkane/Miscellaneous	1.42	0.0457	0.0492
Alcohol/Alkene	1.16	0.1684	0.0619
Cycloalkane/Ketone	0.89	0.0404	0.0807
Alcohol/Ether	0.8	0.026	0.0305
Alkane/Carboxylic Acid	0.36	0.1411	0.145
Ester/Water	0.27	0.0906	0.0719

% - Percentage of mixture type. MAE - Mean Absolute Error in IL-poor and IL-rich phase.

Figure 1, Tables 4 and 5, highlight that the most represented classes of organic molecule-based mixtures in the training set are accurately predicted. Alkane/aromatic, alcohol/alkane, aromatic/cycloalkane, alcohol/water are examples of success. It's important to highlight that mixtures of cycloalkane/cycloalkene, alkane/alkene, alcohol/ester, miscellaneous/water, aromatic/miscellaneous and saccharide/water have been well predicted even considering the relatively low percentage of those mixtures in the training set - Figure 1, Tables 4 and 5. These results highlight the capacity of the optimized models and the MOLMAP approach to encode and predict the phase behaviour profile of different chemical systems (training vs. evaluation sets). Additionally, the results obtained for the saccharide/water mixtures, Table 5 vs. Figure 1, highlight the real possibility for application in the design of aqueous biphasic systems.

Table 5. Distribution of class types in the mixtures of molecules and results of RF predictions for each type (test set).

		IL-poor phase	IL-rich phase
Mixture Type	**%**	**MAE**	**MAE**
Alkane/Aromatic	46.59	0.0305	0.0252
Alcohol/Alkane	7.22	0.0590	0.0394
Alkane/Miscellaneous	6.92	0.0697	0.0531
Saccharide/Water	4.85	0.0215	0.0628
Miscellaneous/Water	3.46	0.0794	0.0881
Alkane/Heteroaromatic	3.26	0.0560	0.0371
Alcohol/Ester	3.17	0.0818	0.0857
Aromatic/Cycloalkane	2.77	0.0409	0.0159
Alcohol/Water	2.37	0.0252	0.0357
Ketone/Water	2.27	0.2377	0.0841
Aromatic/Heteroaromatic	2.27	0.0540	0.0814
Alcohol/Alkene	2.08	0.1274	0.0543
Carboxylic Acid/Water	1.88	0.0165	0.0350
Alkane/Alkene	1.68	0.0878	0.0247
Cycloalkane/Cycloalkene	1.48	0.0626	0.0173
Carboxylic Acid/Ester	1.48	0.0645	0.0890
Alkane/Ester	1.38	0.0691	0.0932
Alcohol/Ether	1.09	0.1182	0.0375
Aromatic/Miscellaneous	0.99	0.0660	0.0535
Alkane/Ketone	0.69	0.0279	0.0187
Alkane/Cycloalkane	0.69	0.1192	0.0139
Alcohol/Ketone	0.59	0.0801	0.0486
Alkane/Carboxylic Acid	0.40	0.1425	0.0823
Ester/Water	0.40	0.1054	0.1029

Figures 2 and 3 represent two examples of chemical systems predicted with success. The system 1-ethyl-3-methylimidazolium methyl sulphate + 2-propanol + ethyl acetate (Figure 2), was studied from chemoinformatic and experimental perspectives in this work, and has been predicted with success. A fact that helps explaining the good agreement between experimental and predicted composition profiles for IL-poor phase, and the slightly higher deviations for IL-rich phase, Figure 2, is the different representativity of experimental composition ranges in the IL-poor and IL-rich training sets, Table 6. The range of experimental compositions with a higher representativity is 0-0.1 in both sets, but the experimental ranges are more equally distributed in the IL-poor phase set than in that of the IL-rich phase . If the predicted molar fraction of at least one organic molecule, is moderately deviated (ester), the error propagates to the calculated molar fraction of the IL, Figure 2.

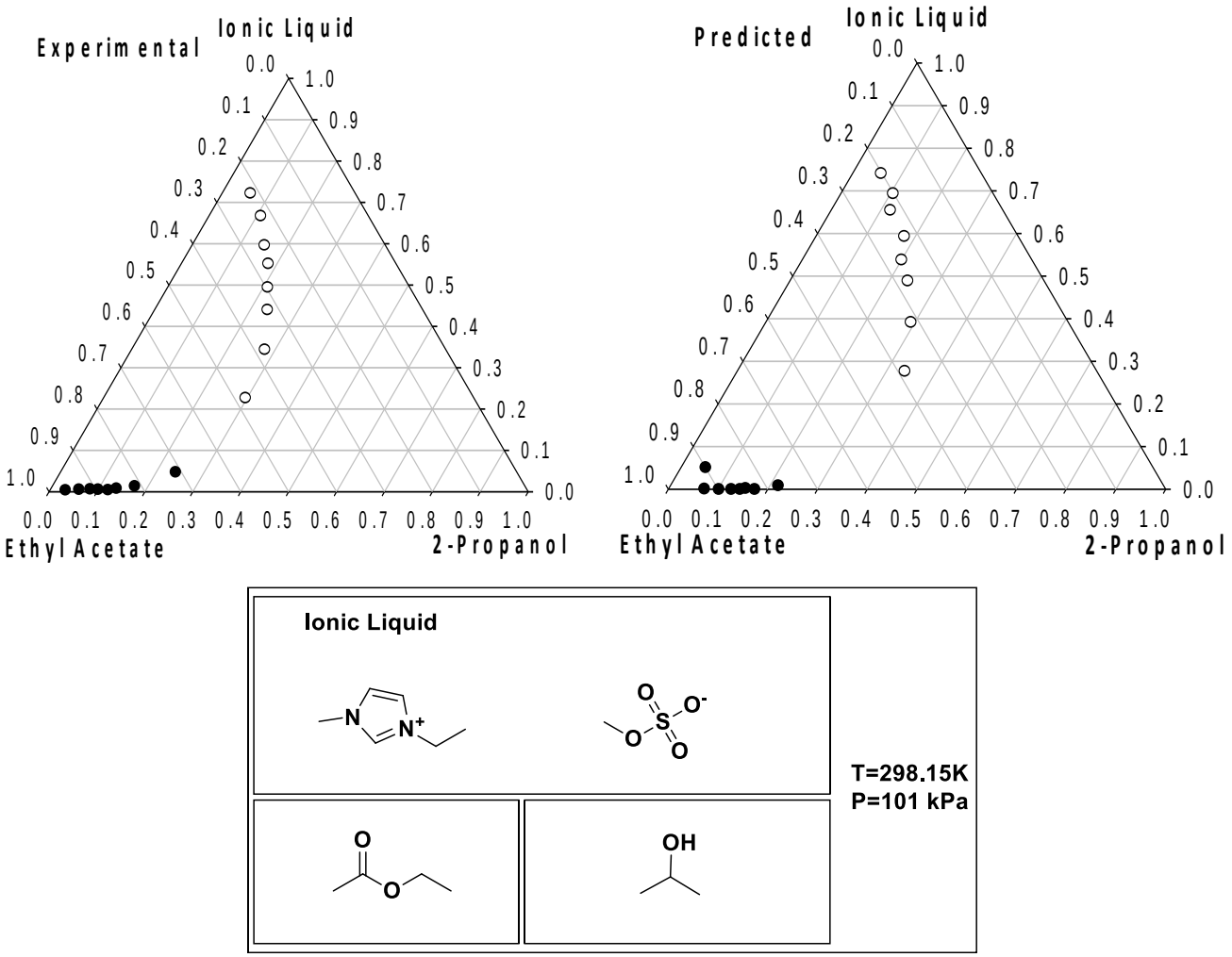


Figure 2. Example of predictive ability of the two models (for IL-poor and IL-rich-phases - black and white dots respectively). Ionic Liquid + Ethyl acetate + 2-Propanol at 298.15 K and 101 kPa as TP conditions. Example of composition profile for alcohol/ester organic molecule mixtures. MAE IL-rich-phase: 2-Propanol - 0.0132, Ethyl Acetate - 0.0426, Ionic Liquid - 0.0419. MAE IL-poor-phase: 2-Propanol - 0.0178, Ethyl Acetate - 0.0265, Ionic Liquid - 0.0158.

Table 6. Distribution of experimental ranges of compositions in the training set . 2 mol - exclusive to molecules. 2 mol + IL - all the compounds (molecules + ionic liquids).

	IL-poor		IL-rich	
	2mol	2mol + IL	2mol	2mol + IL
Range-Exp	%		%	
0-0.1	20.2	46.4	49.2	34.2
0.1-0.2	9.2	6.3	12.9	11.3
0.2-0.3	7.2	4.9	8.9	9.3
0.3-04	6.6	4.4	8.1	8.4
0.4-0.5	7	4.7	6.6	7.4
0.5-0.6	7	4.7	4.8	6.2
0.6-0.7	6.9	4.6	3.7	5.7
0.7-0.8	7.2	4.8	3.3	5.6
0.8-0.9	9.3	6.2	1.9	5.5
0.9-1	19.5	13	0.6	6.2

The 2-propanol/ethyl acetate mixture is used in the extraction of edible oils, and these two organic molecules form an azeotrope that is difficult to break without the use of an entrainer. [35] ILs are a class of compounds that offer multiple advantages such as neglectable vapour pressure, high thermal and chemical stability, high solubility and separation capacity, and are potentially useful for the separation of azeotropes in extractive distillations.

The experimental results obtained for this system (Figure 2) are complementary presented on Table 7, where *x* is the mole fraction. The subscripts 1 and 2 indicate the ethyl acetate and 2-propanol, respectively. The superscripts I and II indicate the IL-poor phase and IL-rich phase, respectively.

Values of solute distribution ratio, β, and selectivity, S, were also calculated using the following equations:

$$\beta = x_2^{II}/x_2^{I} \quad (1)$$

$$S = (x_1^{I}/x_1^{II}) \cdot (x_2^{II}/x_2^{I}) \quad (2)$$

Table 7. Numerical representation of the phase behaviour profile comprising the system ethyl acetate (1), 2-propanol (2) and 1-ethyl-3-methylimidazolium methylsulfate (3).

Ionic liquid poor phase		Ionic liquid rich phase		*β*	*S*
x_1^{I}	x_2^{I}	x_1^{II}	x_2^{II}		
Ethyl acetate (1) + 2-propanol (2) + [C_2MIM] [C_1SO_4] (3)					
0.712	0.240	0.477	0.296	1.23	1.84
0.815	0.172	0.378	0.278	1.62	3.49
0.855	0.137	0.324	0.236	1.72	4.55
0.875	0.121	0.297	0.209	1.73	5.1
0.895	0.099	0.267	0.181	1.82	6.12
0.912	0.082	0.252	0.151	1.84	6.63
0.935	0.060	0.225	0.108	1.81	7.54
0.964	0.032	0.219	0.059	1.84	8.11

In all areas of the phase diagram, the selectivity values are higher than one, which indicates that the extraction of the ethyl acetate from the azeotropic system is possible. The results obtained are comparable to the ones presented by Pereiro et al. [36], where they used the ionic liquid 1-methyl-3-methylimidazolium methylsulfate. As expected, when using ionic liquids with identical anion, but longer alkyl chains attached to the cation, the 2-phase region is smaller. The second case study, Figure 3, comprises the system 2-octylisoquinolin-2-ium isothiocyanate + pyridine + heptane. This example is represented in the optimization set and has been predicted with accuracy. The mixtures alkane/heteroaromatic are considerably represented in the training set, Figure 1, and predicted with accuracy in the optimization and test sets, Tables 4 and 5. Alkane is the most represented class in the training set and heteroaromatic is relatively represented, Figure 4. The ionic liquid, based on the isoquinolinium cation core and the isothiocyanate anion, is poorly represented in the training set, Figures 5 and 6. However the ionic liquid molar fraction is not directly obtained by the IL-poor and IL-rich phase models, but calculated from the predicted molar fractions of pyridine and heptane. The presence and respective proportion of an IL is accounted in the MOLMAP of mixture for a certain datapoint.

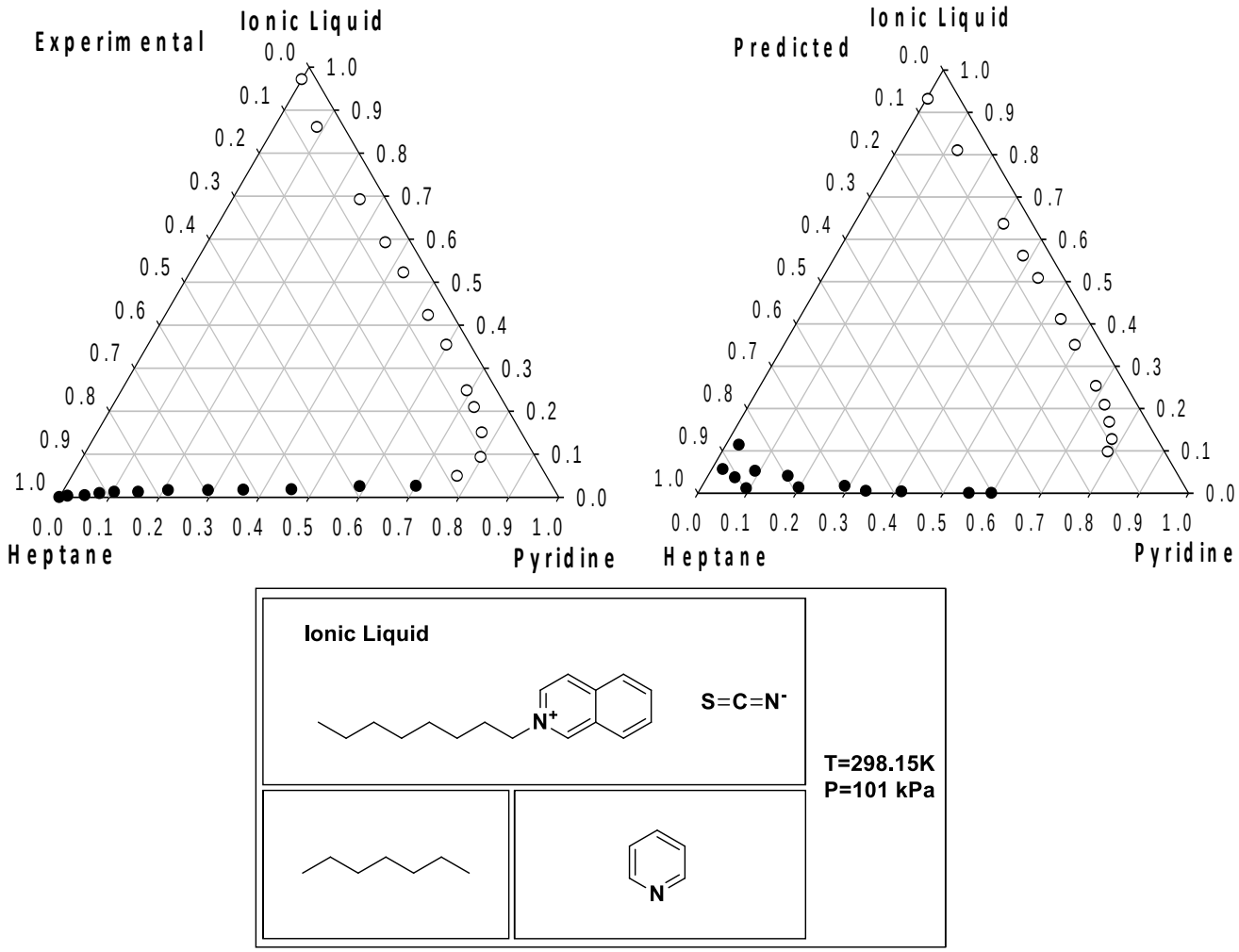


Figure 3. Example of predictive ability of the two models (for IL-poor and IL-rich phases - black and white dots respectively). Ionic Liquid + Heptane + Pyridine at 298.15 K and 101 kPa as TP conditions. Example of composition profile for alkane/heteroaromatic organic molecule mixtures. MAE IL-rich-phase: Heptane - 0.0262, Pyridine - 0.0163, Ionic Liquid - 0.0135. MAE IL-poor-phase: Heptane - 0.0539, Pyridine - 0.0244, Ionic Liquid - 0.0295.

The separation of nitrogen-containing compounds and alkanes find application in petrochemical industry. [37] Once again, the predictions were more accurate for the IL-poor phase, whose training set is more balanced in terms of ranges of molar fractions, Table 6 - 2 mol. This fact helps explaining low/similar values of mean absolute error for IL-poor phase in the optimization and test sets, Tables 8 and 9. The results for the IL- rich phase (optimization and test sets - Tables 8 and 9 respectively - 2mol) present higher differences in mean absolute error between the extreme ranges, molar fractions 0-0.1 and 0.9-1, which are the most and least represented ranges in the training set respectively, Table 6 - 2mol.

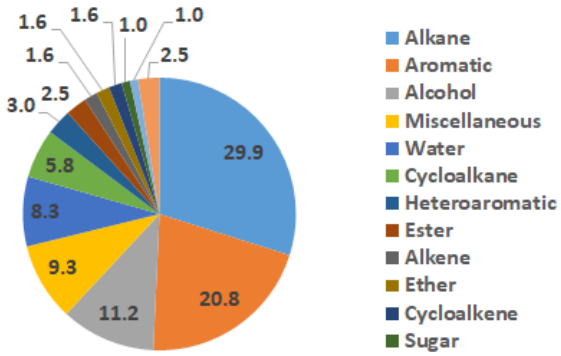


Figure 4. Percentage of organic molecules' families in the training set.

Table 8. Distribution of molar fractions (Range-Exp) in the optimization set, and MAE for the RF predictions.

Optimization Set IL-poor phase				
	2mol		2mol+IL	
Range-Exp	%	MAE	%	MAE
0-0.1	17.01	0.031084	44.43	0.021036
0.1-0.2	9.71	0.049467	6.47	0.045731
0.2-0.3	7.79	0.042727	5.22	0.037092
0.3-0.4	7.79	0.044153	5.40	0.038563
0.4-0.5	8.15	0.045700	5.43	0.038417
0.5-0.6	7.79	0.042188	5.19	0.038236
0.6-0.7	7.84	0.040771	5.22	0.039715
0.7-0.8	7.52	0.039079	5.02	0.041832
0.8-0.9	9.75	0.049642	6.50	0.056028
0.9-1	16.65	0.053566	11.10	0.057147
Optimization Set IL-rich phase				
	2-mol		2-mol+IL	
Range-Exp	%	MAE	%	MAE
0-0.1	53.38	0.015829	36.92	0.016713
0.1-0.2	13.13	0.045987	10.69	0.050028
0.2-0.3	8.64	0.064208	8.25	0.064559
0.3-0.4	7.44	0.050880	7.30	0.059819
0.4-0.5	6.72	0.060336	6.98	0.068385
0.5-0.6	5.12	0.061966	5.79	0.060797
0.6-0.7	2.98	0.068781	5.02	0.057730
0.7-0.8	1.78	0.056969	5.28	0.063578
0.8-0.9	0.62	0.045111	5.25	0.052391
0.9-1	0.18	0.072377	8.52	0.030366

Table 9. Distribution of molar fractions (Range-Exp) in the test set, and MAE for the RF predictions.

Test Set IL-poor phase				
	2 mol		2 mol + IL	
Range-Exp	%	MAE	%	MAE
0-0.1	17.75	0.0420	45.07	0.0298
0.1-0.2	9.35	0.0490	6.33	0.0456
0.2-0.3	7.57	0.0557	5.04	0.0465
0.3-0.4	7.67	0.0516	5.11	0.0415
0.4-0.5	8.01	0.0426	5.34	0.0348
0.5-0.6	7.47	0.0388	4.98	0.0333
0.6-0.7	7.86	0.0440	5.24	0.0456
0.7-0.8	7.42	0.0448	4.95	0.0510
0.8-0.9	9.55	0.0556	6.36	0.0635
0.9-1	17.36	0.0746	11.57	0.0810
Test Set IL-rich phase				
	2 mol		2 mol + IL	
Range-exp	%	MAE	%	MAE
0-0.1	50.84	0.0191	34.55	0.0204
0.1-0.2	13.75	0.0464	11.74	0.0556
0.2-0.3	10.44	0.0635	9.26	0.0661
0.3-0.4	8.31	0.0662	8.21	0.0710
0.4-0.5	5.54	0.0666	6.36	0.0722
0.5-0.6	4.25	0.0710	5.97	0.0777
0.6-0.7	3.21	0.0661	5.41	0.0713
0.7-0.8	2.08	0.0802	5.93	0.0686
0.8-0.9	1.38	0.1114	6.40	0.0536
0.9-1	0.20	0.1782	6.17	0.0347

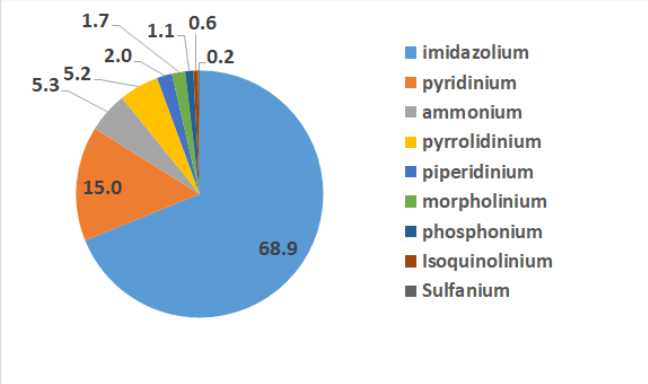


Figure 5. Training set percentage of cation families.

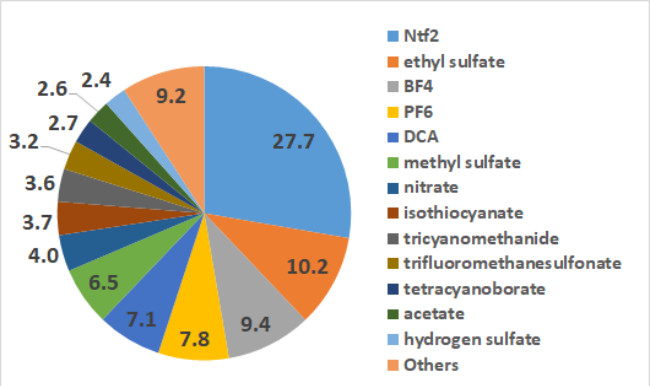


Figure 6. Training set percentage of anions.

It's worth to mention that the optimization set includes two organic molecules, tetrahydrofuran and 1-heptene, 75 datapoints, that are not represented in the training set. The MAE results for their mixtures, are 0.168 and 0.101 for IL-poor and IL-rich phases, respectively. The test set contains six organic molecules, 310 datapoints, not represented in the training set: furan-2-carbaldehyde, tetrahydrofuran, a disaccharide, phthalic acid, bis(tert-butoxy)-ether and 1-heptene. The MAE values for this subset are 0.101 and 0.073 for IL-poor and IL-rich phases, respectively. Ethers, alkenes, sugars and miscellaneous are relatively weakly represented in the training set (Figure 4), but their results are satisfactory. These results illustrate the ability of the MOLMAP encoding method to capture structural features and establish a correct relationship with phase behaviour profile.

This chemoinformatic approach calculates directly the molar fractions of each of the molecules of a mixture. The ionic liquid molar fraction, differently, is obtained by an arithmetic operation involving the two organic molecule's molar fractions, which is faster. However, if an molecule molar fraction is not accurately predicted the error propagates to the ionic liquid.

Conclusion

This work aims at answering a fundamental question: what is the composition profile of a three-compound mixture in LLE, given a certain global composition at given TP conditions?

The chemoinformatic approach here described comprises the use of MOLMAP technology to encode: 1- the global mixture and 2 - the organic molecule to be predicted.

Two complementary models have been built, for IL-rich and IL-poor phases. The models present excellent global predictive accuracy and its performance compares favourably with models trained with fingerprint count descriptors. The validation based on the use of knn algorithm, with k=1, highlights the generalization capacity of MOLMAP/RF approach described in this work.

The MOLMAP encoding approach based on atomic properties, allied to the RF algorithm, enabled models to learn the influence of features present in training set that can be transposed to different structural profiles present in the independent optimization and test sets.

The built models highlight the real possibility to predict the phase behaviour profile of azeotrope-based systems and to assist in the design of aqueous biphasic systems.

Experimental Section

Chemoinformatic Component

Data sets. Three-compound systems in LLE, comprising the property tie-line (the composition of the two phases in equilibrium), have been gathered from ILThermo database. [30] This procedure has been accomplished using in-house developed graphical user interface based on the pyILT2 package. [38]

Weight fractions have been converted into molar fractions.

The training set comprises 6951 tielines. The optimization and test sets are composed by 1123 and 1011 tielines, respectively. Any combination of IL + organic molecule 1 + organic molecule 2 in the optimization and test sets is different from any combination in the training set. The chemical systems contained in the test set comprise combinations of four components that are not contemplated in the training and optimization sets. This assures a correct evaluation of predictive ability.

The sum of the three-component molar fractions of a given phase in equilibria is 1, as expected. This has been normalized to 2 in order to consider the molar fraction of the cation and anion of the IL separately.

The next step comprises the generation of five random points within each tie-line, which was performed with the program Mathematica. [39] This step led to the generation 6951 x 5 = 34755 training set datapoints, 5615 optimization and 5055 test set datapoints.

As the molar fraction of one compound (out of three) can be calculated from the molar fraction of the remaining two, the molar fraction of the ionic liquid was calculated from the molar fraction of the two organic compounds. This approach yielded 34755 x 2 = 69510 datapoints for the training set and 11230 and 10110 datapoints, respectively, to the optimization and test sets.

Each of the two phases in equilibrium comprises a phase that is more IL-rich than the other phase regarding its molar fraction. Considering it, data

sets were prepared separately for the ionic liquid-rich phase and the ionic liquid-poor phase.

The training set temperature range is 278.15-373.9 K. The range of pressures is 1-11343 kPa. The optimization set comprises a range of temperatures of 283.15-348.2 K. The pressure range is 1-101 kPa. Test set range of temperatures is 278.15-337.15 K. The respective pressure range is 81.5-101 kPa.

All the chemical structures of all datasets are standardized, using the standardizer application from Chemaxon. [40] The standardizer operations comprise: 1) mesomerize, 2) add explicit H and 3) clean 3D.

The different datasets were deposited in a public repository. [41]

Representation of molecules and mixtures. Each datapoint of a given dataset is represented by: 1 - numerical descriptors of the four-component mixture (cation + anion + molecule 1 + molecule 2) weighted by the respective four-coordinate random point-based global molar fractions (one out of five random points), 2 - numerical descriptors of the organic molecule (1 or 2) weighted by the respective molar fraction (the respective coordinate of the random point-based global molar fractions), 3 - temperature and pressure (Figure 7). The sum of the four-coordinates (random point) is normalized to 2.

The Molecular Maps of Atom-level Properties (MOLMAPs) [18] were used to encode the mixture and the molecule to be predicted. The MOLMAP approach is based on the atomic pattern of activation of a given chemical system in a *Kohonen* neural network, Java Tools for Neural networks (JATOON), built in our labs, implements this algorithm. [42,43]

Nine atomic properties/features were calculated with the program cxcalc [40]: 1- atomic number, 2- pi charge, 3- sigma charge, 4- total charge, 5- orbital electronegativity sigma, 6- polarizability, 7- hindrance, 8- hydrogen acceptor site and 9- hydrogen donor site.

Each of these nine atomic properties/features are normalized from 0-1 using equation 3:

Normalized prop.=(prop.- minimum)/(maximum-minimum) (3)

The maximum and minimum values for that specific property/feature were retrieved from the training set.

The next step comprised the random selection of 7500 atoms of anions, 7500 atoms of cations and 14997 atoms of molecules 1 and 2 from the training set. All the 29997 atoms (characterized by the nine normalized properties/features) were used to train a *Kohonen* neural network (size 30x30 or 25x25). This self-organising map was then used to encode the mixtures (Figure 7).

A different *Kohonen* neural network (size 20x20 or 25x25) was trained with 29998 randomly-selected atoms of molecules 1 and 2 from the training set. This network is used to obtain numerical descriptors of a given molecule (molecule 1 or 2), a datapoint (Figure 7). To encode a molecule, all its atoms (represented by its nine properties) are submitted to the trained network and the MOLMAP representation of the molecule is the pattern of the activated neurons by the atoms of the molecule. Optionally the activation may extend to the neighbors of the winning neuron, usually with a progressively lower value of activation as the distance to the winning neuron increases. In the optimized models, the neighbour neurons until the 3rd neighbourhood (n1-n2-n3) were also activated; the values of w-n1-n2-n3 activation were 1-0.75-0.5-0.25.

Similarly, a mixture is encoded by submitting all the atoms of its components to the trained self-organising map.

The activation pattern of the 2D map is converted into a vector by concatenation of lines. Figure 7 illustrates the generation of a MOLMAP for a mixture, with activation values displayed inside the activated neurons (for simplicity, in this figure the activation was restricted to the winning neurons). The values resulting from activations are weighted according to the global molar fraction of the component.

The *Kohonen*-based MOLMAP has a toroidal topology*, i.e.,* the neurons at the right end are direct neighbors of the neurons at the left end and the same for neurons represented at the top and bottom.

Two models have been built, based on this form of input codification, one to predict the molar fraction of the organic molecule (1 or 2) in the ionic liquid rich phase, and the other, similarly, for the ionic liquid poor phase. The IL composition in each of the phases in equilibrium is obtained by a simple arithmetic operation involving organic molecule 1 and 2 molar fractions on the phase being evaluated.

Machine learning. Two models were built, one to predict the profile of the ionic liquid-rich phase and the other for the ionic liquid-poor phase. Each model provides as output the molar fraction of the organic molecule (1 or 2) of a given phase in equilibrium.

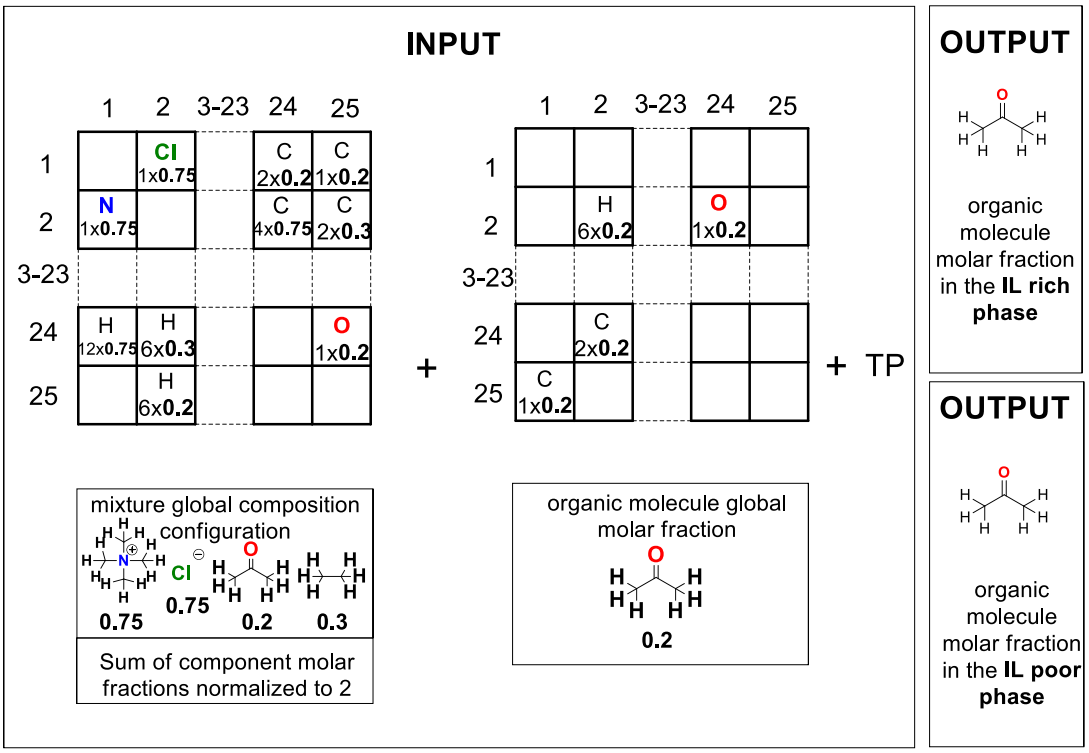


Figure 7. The MOLMAP of the mixture results from the activation of neurons on the 2D surface (grid) of a Kohonen map by the atoms of the mixture, multiplied by the molar fractions of the components they belong (left). The MOLMAP of the organic molecule is generated in a similar way. Both are used as independent variables, additionally to the temperature and pressure (INPUT). The dependent variable is the molar fraction of the organic molecule (1 or 2) in the IL-rich and IL-poor phase

The IL-rich and IL-poor phase models were obtained with the Random Forest algorithm, [28], version 4.6-14, using R platform (version 3.5.3), [29]. Each model included 1000 trees, and the number of input descriptors randomly selected to be evaluated at each tree node (*mtry)* is 200.

The models were evaluated on the basis of the squared Pearson correlation coefficient (R^2), mean absolute error (MAE) and root mean square (RMS) for the obtained predictions.

Optimization. The representation of the molecules and mixtures was optimized by training models with a subset of the training set (tr1 - 20030 out of 69510 datapoints) and evaluating them with the optimization set. The optimized parameters consist in:

- Size of MOLMAPs, for :

  a) the mixture representation (30x30 - 900 descriptors or 25x25 - 625 descriptors).

  b) the single organic molecule representation (molecule 1 or molecule 2) present in a given chemical system of four components, (20x20 - 400 descriptors or 25x25 - 625 descriptors), for which the phase behaviour profile is predicted.

- MOLMAP pattern of activation for each atom considering the winning neuron and three-level neighbourhood (w-n1-n2-n3): 1-0-0-0, 1-0.5-0-0, 1-0.66-0.33-0 and 1-0.75-0.5-0.25.

The optimized parameters selected to build the final ionic liquid poor and rich models consist on the use of:

i) MOLMAPs of mixture with dimensionality 25x25 and generical atomic pattern of activation 1-075-05-025.

ii) MOLMAPs of organic molecule being evaluated 25x25 with an atomic pattern of activation of 1-075-05-025.

The optimum MOLMAP-based parameters were used to obtain the two final models (for IL poor and IL rich phases) built with the complete training sets of 69510 datapoints. They were evaluated with the optimization and independent test sets.

Laboratorial Component

Materials: 1-ethyl-3-methylimidazolium methylsulphate was acquired from IOLITEC with a stated purity higher than 98% and was kept under vacuum at 0.1 Pa and 323 K for at least 24 h, before preparing any mixture. The water content was always below 200 ppm. Ethyl acetate and 2-Propanol were supplied by Carlo Erba and are anhydrous solvents with purity over 99.8% and water content below 100 ppm and 300 ppm, respectively.

Method: The phase diagram of the system (1-ethyl-3-methylimidazolium methylsulphate + 2-propanol + ethyl acetate) has been studied experimentally to complement the data already available in literature. The determination of each tie-line was performed experimentally using NMR to obtain the composition of each of the phases in equilibria. The validity of the methodology used was confirmed with the composition of two mixtures that were prepared with known compositions of the three components. For the determination of the equilibrium composition of each phase at 298.15 K, known weights of the three components were mixed in a mixture composition that falls in the 2-phase region of the phase diagram. The mixture was vigorously stirred and was left to separate in a glass separatory funnel for at least 24h. The top and the bottom phases were

separated and weighted. NMR analysis using an appropriate deuterated solvent (D_2O) allowed the access to the composition of each phase.

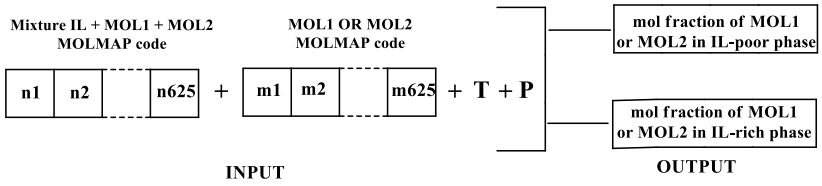
Acknowledgements

We thank Portuguese Foundation for Science and Technology Project: PTDC/EQU-EQU/30060/2017. This work was supported by the Associate Laboratory for Green Chemistry - LAQV which is financed by National Funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020).

Keywords: Ionic Liquid • Phase Behaviour • Chemoinformatics • Codification • Big Data

[1] L. Zhou, S. Pana, J. Wang, A. V. Vasilakos, *Neurocomputing* **2017**, 237, 350-361.
[2] T. Engel, J. Gasteiger, *Chemoinformatics*, Wiley-VCH, Weinheim, **2018**, pp 465-495.
[3] J. M. Winnert, V. K. P. J. Devi, J. F. Brennecke, *Ind. Eng. Chem. Res.* **2019**, *58*, 22633-22639.
[4] P. Wang, D. Xua, P. Yan, J. Gao, L. Zhang, Y. Wang, *J. Mol. Liq.* **2018**, *261*, 89-95.
[5] J. Dhanalakshmi, P. S. T. Sai, A. R. Balakrishnan, *Ind. Eng. Chem. Res.* **2013**, *52*, 16396-16405.
[6] X. Liu, D. Xua, B. Diao, L. Zhang, J. Gao, D. Liu, Y. Wang, *J. Mol. Liq.* **2019**, *275,* 347-353.
[7] P. Jia, Z. Zhao, X. Zhang and C. Hao, *J Chem Technol. Biotechnol.* **2021**, *96*, 64-77.
[8] F. Shaahmadi, M. A. Anbaz, *J. Mol. Liq.* **2017**, *232*, 396-407.
[9] M. G. Freire, L. M.N.B.F. Santos, I. M. Marrucho, J. A.P. Coutinho, *Fluid Phase Equilib.* **2007**, *255*, 167-178.
[10] S. Gao, X. Chen, R. Abro, A. A. Abdeltawab, S. S. Al-Deyab, G. Yu, *Ind. Eng. Chem. Res.* **2015**, *54*, 9421-9430.
[11] M. A. Rocha, Y. Zhang, E. J. Maginn, M. B. Shiflett, *J. Chem. Phys.* **2018**, *149*, 164503.
[12] J. Jacquemin, M. Bendová, Z. Sedláková, M. Blesic, J. D. Holbrey, C. L. Mullan, T. G. A. Youngs, L. Pison, Z. Wagner, K. Aim, M. F. C. Gomes, C. Hardacre, *ChemPhysChem* **2012**, *13*, 1825-1835.
[13] Y. Lyu, J. F. Brennecke, M. A. Stadtherr, *Ind. Eng. Chem. Res.* **2020**, *59*, 8871-8893.
[14] A. Fazlali, P. Koranian, R. Beigzadeh, M. Rahimi, *Korean J. Chem. Eng.* **2013**, *30*, 1681-1686.
[15] P. A. Robles, T. A. Graber, M. Aznar, *Fluid Phase Equilib.* **2010**, *296*, 154-158.
[16] F. Sarlak, T. Pirhoushyaran, F. Shaahmadi, Z. Yaghoubi, B. Bazooyar, *Sep. Sci. Technol.* **2018**, *53*, 2935-2951.
[17] L. D. Simoni, A. Chapeaux, J. F. Brennecke, M. A. Stadtherr, *Ind. Eng. Chem. Res.* **2009**, *48*, 7257-7265.
[18] Q.-Y. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2005**, *45*, 1775-1783.
[19] B. Hemmateenejad, A. R. Mehdipour, P. L. A. Popelier, *Chem. Biol. Drug Des.* **2008**, 72, 551-563.
[20] G. V. S. M. Carrera, S. Gupta, J. Aires-de-Sousa, *J. Comput. Aided Mol. Des.* **2009**, *23*, 419-429.
[21] D. A. R. S. Latino, Q-Y. Zhang, J. Aires-de-Sousa, *Bioinformatics* **2008**, *24*, 2236-2244.
[22] D. A. R. S. Latino, J. Aires-de-Sousa, *Angewandte, Chem. Int. Ed.* **2006**, *45*, 2066-2069.
[23] Q-Y. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2007**. 47, 1-8.
[24] S. Gupta, S. Matthew, P. M. Abreu, J. Aires-de-Sousa, *Bioorg. Med. Chem.* **2006**, *14*, 1199-1206.
[25] G. V. S. M. Carrera, M. Nunes da Ponte, L. P. N. Rebelo, *ChemPhysChem* **2019**, *20*, 2767-2773.
[26] G. V. S. M. Carrera, M. Nunes da Ponte, *Chemistry-Methods* **2021**, *1*, 214-223.
[27] G. V. S. M. Carrera, J. Inês, C. E. S. Bernardes, K. Klimenko, K. Shimizu, J. N. Canongia Lopes, *ChemPhysChem* **2021**, *22*, 1-12.
[28] L. Breiman, *Machine Learning* **2001**, *45*, 5-32.
[29] https://cran.r-project.org/ (last access July 2022).
[30] https://ilthermo.boulder.nist.gov/. (last access July 2022).
[31] P. E. Tetlock, D. Gardner, *Superforecasting: the art and science of prediction*, Broadway Books, New York, **2015**.
[32] http://www.yapcwsoft.com/dd/padeldescriptor/. (last access July 2022).
[33] I. Wallach, A. Heifets, *J. Chem. Inf. Model.* **2018**, *58*, 916-932.
[34] https://scikit-learn.org/.(last access July 2022).
[35] Z. Zhang, A. Zhang, K. Wu, Q. Zhang, A. Hu, W. Li, *Fluid Phase Equilib.* **2016**, *429*, 331-338.
[36] A. B. Pereiro, A. Rodríguez, *J. Chem. Therm.* **2007**, *39*, 1608-1613.
[37] K. Kedra-Królik, M. Fabrice, J-N. Jaubert, *Ind. Eng. Chem. Res.* **2011**, *50*, 2296-2306.
[38] Q. Dong, C. D. Muzny, A. Kazakov, V. Diky, J. W. Magee, J. A. Widegren, R. D. Chirico, K. N. Marsh, M. Frenkel, *J. Chem. Eng. Data* **2007**, *52*, 1151-1159.
[39] https:/www.wolfram.com/mathematica/. (last access July 2022)
[40] JChemSuite was used for standardization of chemical structures (standardizer) and calculation of atomic properties (cxcalc), JChemSuite version 19.4.0, ChemAxon (https://www.chemaxon.com) (last access July 2022)
[41] G. V. S. M. Carrera, M. L. Cruz, K. Klimenko, J. M. S. S. Esperança, J. Aires-de-Sousa, Phase behaviour profile of three compound mixtures in Liquid-Liquid Equilibrium.
figshare. https://dx.doi.org/10.6084/m9.figshare.20402679
[42] http://neural.dq.fct.unl.pt/jas/jatoon/ (last access July 2022).
[43] J. Aires-de-Sousa, *Chemom. Intell. Lab. Syst*. **2002,** *61* 167-173.

Entry for the Table of Contents



The MOLMAP encoding method, based on the atomic pattern of activation of a given chemical system on a Kohonen neural net was used to represent: a- 3-compound mixture (Ionic Liquid + molecule1 + molecule 2) and b- molecule1 or molecule2 of that mixture at certain temperature and pressure. The Random Forest algorithm predicts its phase behavior profile.