FULL PAPER

The Solubility of Gases in Ionic Liquids: A Chemoinformatic Predictive and Interpretable Approach

Gonçalo V. S. M. Carrera,[a]* João Inês,[a] Carlos E. S. Bernardes,[b] Kyrylo Klimenko,[a] Karina Shimizu,[c] José N. Canongia Lopes[c]

[a] Dr. G. V. S. M., Carrera*
Msc. J. Inês
Dr. Kyrylo Klimenko
Chemistry Department
LAQV-REQUIMTE, NOVA School of Science and Technology
2829-516 Caparica, Portugal
goncalo.carrera@fct.unl.pt

[b] Dr. C. E. S. Bernardes
Centro de Química Estrutural,
Faculdade de Ciências, Universidade de Lisboa
1749-016 Lisboa, Portugal

[c] Prof. Dr. J. N. Canongia Lopes
Dr. K. Shimizu
Centro de Química Estrutural,
Department of Chemical and Biological Engineering,
Instituto Superior Técnico, Universidade de Lisboa,
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

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

Abstract: This work comprises the study of solubilities of gases in ionic liquids (ILs) using a chemoinformatic approach. It is based on the codification, of the atomic inter-component interactions, cation/gas and anion/gas, which are used to obtain a pattern of activation in a *Kohonen* Neural Network (MOLMAP descriptors). A robust predictive model has been obtained with the Random Forest algorithm and used the maximum proximity as a confidence measure of a given chemical system compared to the training set. The encoding method has been validated with molecular dynamics. This encoding approach is a valuable estimator of attractive/repulsive interactions of a generical chemical system IL+Gas. This method has been used as a fast/visual form of identification of the reasons behind the differences observed between the solubility of CO_2 and O_2 in BMIM.PF_6 at identical temperature and pressure (TP) conditions, The effect of variable cation and anion effect has been evaluated.

Introduction

Ionic Liquids (ILs) have been studied in the last decades, from its fundamentals to applications. The possibility of simply combining a cation with an anion provides new perspectives in order to tune physico-chemical properties/activities and structural organization. There are virtually infinite combinations between cation and anion, which convert the proper choice by trial and error for a given application an impossible task. The recent use of machine-learning techniques and the access to databases with high content of information enhanced the implementation of ILs. [1,2] Fundamental properties such as melting points, [3,4] solubilities, [5-7] and viscosities [8,9] have been modelled. Due to its characteristic nature, many ILs present different domains, such as zones dominated by coulombic atractions/repulsions and other zones dominated by van der Waals interactions. ILs complex behaviour and versatility stems from its structural heterogeneity and dynamic behaviour. Gas solubility is an important characteristic of ILs in the context of gas separation, [10] optimization of reactions, [11,12] greenhouse gas capture, [13] as antisolvent [14] and absorption of refrigerant gases. [15] The gas solubility in ILs depends on conditions of pressure, temperature, attractive/repulsive interactions and free volume within the IL. Predictive models based on UNIFAC group contribution, COSMO, Equations of State, and machine learning approaches are reported in the literature. [5,16] However, limited number of ILs and gases tested, difficulty of interpretation, and relatively poor predictive ability characterize these approaches. The use of chemoinformatic machine-learning approaches offers the possibility of finding hidden relationships on high-information content databases and fast prediction for new situations. The Random Forest (RF) machine-learning algorithm, tested in this work, has been used with success in multiple situations. Some examples comprise its use on solar cell design, [17] chromatographic retention indices prediction, [18] applicability domain modelling, [19] and prediction of aquatic toxicity. [20]

The form of codification is important in the context of finding relevant features in an IL-based system, providing predictive and interpretation abilities. [9,21] The Molecular Maps of Atom-level Properties (MOLMAPs), previously developed and tested, permit such approach by mapping structural features of a system in a fixed-dimension's *Kohonen* network, according its property's profile. This characteristic permit to compare, in a straightforward form, systems of different nature/number of components. MOLMAPs have been tested first on classification of chemical reactions without assignment of reaction center. [22] Moreover, and considering that most reaction databases contain only examples of reactions where reaction occurs, the MOLMAP concept permits to attain the code of a pseudo-compound that does not react by comparison between MOLMAPs of products and reactants. [23] Different approaches illustrate the flexibility of this concept. [24-26]

The MOLMAP codification system is a generical form of finding similarities/differences within structural moieties composing a chemical framework. This codification technology has been tested to encode atomic [9] and bond profiles [23]. This work expands the concept of MOLMAPs to inter-component interactions, namely all combinations of pairs of atoms between cation and gas / anion and gas in a binary system IL/gas.
Different approaches are reported in the literature regarding the use of group contribution fragments in order to predict solubilities. [5] Similarly, this approach is based on fragments, in concrete, inter-component interactions within pairs of atoms of gas/anion and gas/cation. The main advantage is the generation, in an automated form, of numerical descriptors, based on those binary/atomic intercomponent interactions. It is also worth to mention that this form of codification establishes similarities and differences of each combined intercomponent ion/gas interaction, (based on each respective interaction property's profile), characterizing a given data point-chemical IL + gas system. This is a generical form of grouping fragments and, by the average within the seven pairs of combined inter-component atomic measures (pair ion/gas), verify whether an interaction is attractive or repulsive.

The present work comprises the combined use of this new form of codification, based on MOLMAP methodology, temperature and pressure, in order to characterize a given chemical system, and the RF algorithm, [27,28] in order to find a straightforward relationship between an ILs system's configuration and the property of interest, the molar gas solubility. This form of codification permits to identify whether an inter-component interaction ion/gas is attractive or repulsive Finally, the interactions profile found using the MOLMAP approach is confronted with predictions obtained using molecular dynamics (MD) simulation results.

Results and Discussion

This work comprises the test, for the first time, of the MOLMAP encoding system, based on intercomponent interactions ion/gas, in combination with the RF algorithm, in order to predict and interpret the reasons for a given gas solubility profile in a generical IL.

Different approaches are reported in the literature. Jaschik et. al [29] tested COSMO-SAC methodology in order to predict the solubility of a restricted number of gases in ILs. The most consistent results are obtained for CO_2. Manan et. al. [30] used COSMOthermX platform to qualitatively predict the solubility of an extended number of gases in 27 ILs. Yokozeki and Shiflett [31] used a modified Van der Walls equation of state to correlate and predict accurately the solubility for a restricted number chemical systems, including CO_2, SO_2, NH_3 and a hydrofluorocarbon in imidazolium-based ionic liquids.

Other relevant models have been developed, however with a restricted applicability: Venkatraman and Alsberg [32] tested decision trees and Random Forest algorithms in order to satisfactorily predict the CO_2 solubility in 185 ionic liquids at different conditions of temperature and pressure. Baghban et. al. [33] used multi-layer perceptron artificial neural network, Peng-Robinson, and Soave-Redlich-Kwong equations of state in order to estimate the CO_2 solubility in ILs at different conditions of pressure and temperature. Shafiei et. al. [34] applied artificial neural networks in order to correlate and estimate the H_2S solubility in ILs with variable temperature and pressures associated.

The approach here reported, besides the generation of predictive models in a quantitative form, permits the interpretation, within a given IL + gas system, whether an interaction is attractive or repulsive, explaining a given solubility profile.
This is a generalizable approach, and differently from previous studies, includes the solubilities of different gases in multiple ionic liquids at different conditions of temperature and pressure.

The first step of this approach consists on finding a robust predictive model (Table 1). It comprises the tuning of the MOLMAP activation pattern, considering the winning neuron, (the position of the *Kohonen* neural network activated by a given inter-component interaction), and the respective neighbourhood, up to three levels of distance (check Methods section).
The Random Forest algorithm *mtry* parameter has been additionally tuned (Table 1).

The test sets 1 and 2 comprise systems where the combination (cation, anion and gas) is different to all the combinations present in each system of the training set. In most situations, each combination cation, anion and gas comprise diverse data points at different temperature and/or pressures. The use of these two independent test sets is a stricter form of evaluating the predictive ability as the structural motifs, temperature and pressure change. The OOB predictions do not ensure structural difference. The results in Table 1 reveal that the MOLMAP activation pattern 1-0.75-0.5-0.25 and the *mtry* value of 600 ensure the best results.

Table 1. Model optimization: The effect of 1 - 30x30 MOLMAP pattern of activation 1-0-0-0, 1-05-0-0, 1-066-033-0, 1-0.75-05-025 as *Kohonen*-Neural Network winning neuron-1st-2nd-3rd level neighbourhood activation by each intercomponent interaction gas/ion of a given chemical system datapoint, 30x30 MOLMAP is equivalent to 900 descriptor's vector, after concatenation of the lines of the *Kohonen* matrix.

					1-0-0-0				
		mtry=200			mtry=400			mtry=600	
	R^2	MAE	RMS	R^2	MAE	RMS	R^2	MAE	RMS
TR	0.959	0.024	0.0399	0.9889	0.0115	0.0206	0.9942	0.0082	0.014633
OOB	0.933	0.03	0.0499	0.9697	0.0188	0.0333	0.9768	0.0162	0.028871
TE1	0.729	0.062	0.0933	0.7135	0.0579	0.0963	0.6992	0.0582	0.101288
TE2	0.419	0.084	0.1664	0.4025	0.0817	0.169	0.3999	0.0812	0.168644
					1-05-0-0				
		mtry=200			mtry=400			mtry=600	
	R^2	MAE	RMS	R^2	MAE	RMS	R^2	MAE	RMS
TR	0.958	0.0249	0.0404	0.9893	0.0004	0.0203	0.9946	0.0078	0.014227
OOB	0.9292	0.0312	0.0513	0.9698	0.0011	0.0334	0.9783	0.0155	0.028071
TE1	0.7759	0.0601	0.0865	0.7754	0.0526	0.0869	0.8016	0.0528	0.084003
TE2	0.5381	0.086	0.1468	0.5816	0.0771	0.1392	0.5992	0.0743	0.136498
					1-066-033-0				
		mtry=200			mtry=400			mtry=600	
	R^2	MAE	RMS	R^2	MAE	RMS	R^2	MAE	RMS
TR	0.9686	0.025	0.0402	0.99	0.0112	0.0197	0.995	0.0076	0.013756
OOB	0.9296	0.0314	0.0512	0.9708	0.0186	0.0329	0.9792	0.0153	0.027551
TE1	0.8738	0.0533	0.0734	0.9055	0.0398	0.0572	0.9048	0.0374	0.056188
TE2	0.4925	0.0858	0.1548	0.5118	0.0774	0.1517	0.448	0.0814	0.160385
					1-075-05 025				
		mtry=200			mtry=400			**mtry=600**	
	R^2	MAE	RMS	R^2	MAE	RMS	**R^2**	**MAE**	**RMS**
TR	0.9579	0.0254	0.0405	0.9894	0.0116	0.0203	**0.9947**	**0.0078**	**0.014194**
OOB	0.9285	0.0319	0.0516	0.9694	0.019	0.0336	**0.9781**	**0.0154**	**0.028192**
TE1	0.8574	0.0584	0.0799	0.8722	0.0476	0.067	**0.8631**	**0.0444**	**0.066503**
TE2	0.6265	0.077	0.1355	0.6831	0.0663	0.1254	**0.7**	**0.063**	**0.121748**

Random Forest algorithm *mtry* parameter. R^2- Square of Pearson Correlation Coefficient, MAE - Mean Absolute Error, RMS - Root Mean Square Error, TR - Training Set, OOB - Training Set Out of Bag Predictions, TE1 - Independent Test Set 1, TE2 - Independent Test Set 2.

The more robust model has been evaluated further (MOLMAPs 30x30 dimension 1-075-05-025 pattern of activation, Random Forest control parameter *mtry* = 600. The two independent test sets (Tables 2 and 3) have been used for that end. These sets comprise systems where the combination of cation, anion and gas is different from any combination present in the training set.

Table 2. Test Set 1, the effect of the Random Forest proximity threshold respective to the Training set as confidence measure to evaluate predictions.

R^2	RMS	MAE	%	N-datapoints	Max-Prox-Threshold
0.8631	0.066503	0.044433	100	702	>=0.041
0.8643	0.067058	0.044464	95	666	>0.1
0.8767	0.064318	0.041367	84	590	>0.15
0.9126	0.05445	0.035307	72	502	>0.2
0.9374	0.046936	0.031062	61	427	>0.25
0.9624	0.038445	0.02575	47	333	>0.3
0.9735	0.033629	0.023356	34	236	>0.35

R^2: Square of Pearson Coefficient, RMS: Root Mean Square error, MAE: Mean Absolute Error

The results for the complete Test Set 1, reported in Tables 1 and 2, indicate a very good correlation and low deviations between experimental and the respective predictive values. The results become excellent with the increment of threshold maximum proximity respective to the training set. This maximum proximity measure (between a generical datapoint in the Test Set 1 and the most resembling datapoint in the training set) is an indication of confidence in a given prediction (Table 2).
The prediction results for the complete Test Set 2 are good, however if we consider a progressive increment on the proximity measure threshold (similarly to Test Set 1), the results become excellent (Table 3).

R^2	RMS	MAE	%	N-datapoints	Max-Prox-Threshold
0.7	0.121748	0.063029	100	635	>=0.085
0.7141	0.124325	0.062256	92	584	>0.15
0.7267	0.126055	0.061538	82	518	>0.2
0.7281	0.12488	0.06046	68	434	>0.25
0.8323	0.090612	0.039916	53	338	>0.3
0.8791	0.072242	0.02894	42	265	>0.35
0.9426	0.046085	0.018643	36	226	>0.4
0.9642	0.036399	0.014223	31	197	>0.45
0.9853	0.024198	0.00976	28	175	>0.5

Table 3. Effect of the threshold proximity of Test set 2 systems respective to training set.

R^2: Square of Pearson Coefficient, RMS: Root Mean Square error, MAE: Mean Absolute Error.

The Test Set 2 comprises 27 chemical systems (each system contains different TP datapoints). Four of those chemical systems comprise highly skewed data points when comparing experimental and predicted solubility measures (Figure 1 - Systems a-d).

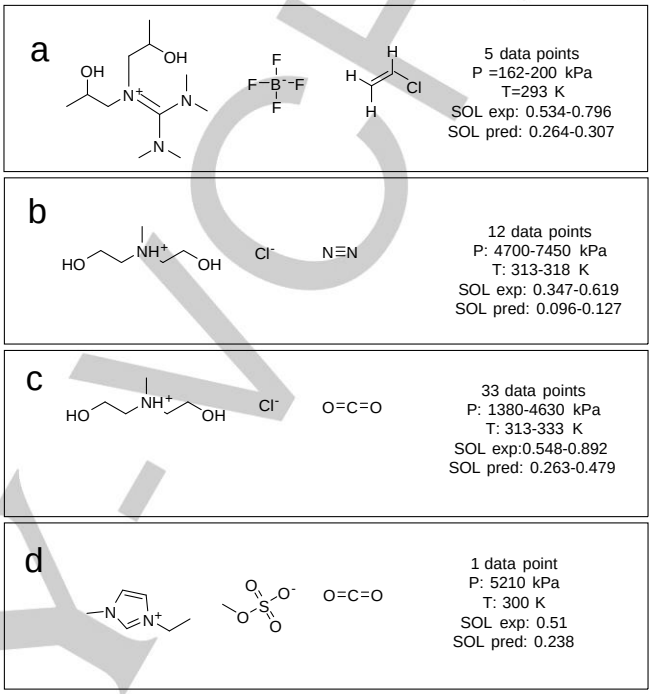


Figure 1. Test set 2 with 4/(27) structural systems predicted with lower accuracy.

The cation in system a (test set 2), is not represented in the training set. This situation contributes to low predictive ability. The IL in systems b and c of test set 2 is also represented in the training set, however with oxygen as dissolved gas. One possible reason for the poor results for these two systems is the high importance of gases in the codification of MOLMAPs, as the inter-component interactions correspond to cation/gas and anion/gas atomic combinations. In addition, the experimental error usually associated to solubility measures for oxygen and nitrogen is high. These two factors may contribute synergistically for the obtained error measures. If we remove these four systems (51 data points), the predictions for the test become very accurate (R^2: 0.879, MAE: 0.0351, RMS: 0.0554), corresponding to approximately 92% of the complete test set 2.

A different approach, regarding test set 2, comprises the study of the influence of the degree of difference respective to the training set (Table 4). If a cation, existent in the test set 2, is not available in the training set, the correlation and accuracy of the model is poor - 49 data points. Differently, if a gas in the test set 2 is not present in training set the accuracy is very good, 22 data points. Similarly, when an IL in the test set 2 is not represented in the training set the results are very accurate, 214 data points.
These two last results represent a certain capacity for extrapolation by the model and encoding method.

Table 4. Statistical characterization of test set 2 concerning the degree of difference respective to training set systems.

Different Cation			
R^2	MAE	RMS	N data points
0.46	0.087	0.139	49
Different Gas			
0.87	0.025	0.027	22
Different Combination IL			
0.84	0.034	0.076	214

R^2: Square of Pearson Coefficient, RMS: Root Mean Square error, MAE: Mean Absolute Error.

The more robust model has been validated by randomization of the solubilities in the training set (5x) and prediction for both test sets. The correlation in the five assays, for both sets, is nearly 0. The RMS maximum and minimum for test set 1 is [0.172-0.187] and [0.215-0.226] for test set 2. The values of MAE are [0.136-151] for test set 1 and [0.156-0.166] for test set 2. These results show that there's an intrinsic order between the numerical description of a given system and the gas solubility value when we compare the chosen model with the randomized models. The more robust model has been additionally validated by 10x cross validation, where for each turn, 9/(10) subsets are used to build a model, and the set aside subset is used to obtain predictions. The procedure is carried out 10 times. Each time a different subset is used to obtain predictions. Each of the 10 subsets is different from the remaining 9 as the combination of cation, anion and gas in that subset is different from the combination's existent in the remaining subsets. The R^2, comprising all the ten subsets set aside for prediction is 0.77, RMS of 0.090 and MAE of 0.055.

The codification system, as stated before, is based on a pattern of activation in a *Kohonen* neural network of the inter-component interactions regarding all the combinations of pairs of atoms cation/gas and anion/gas within a given chemical system IL + gas. Each inter-component interaction is represented by seven combined properties. 1: atomic pi charge of ion x atomic pi charge of gas, 2: atomic sigma charge of ion x atomic sigma charge of gas, 3: atomic total charge of ion x atomic total charge of gas, 4: abs(atomic orbital electronegativity sigma of ion - atomic orbital electronegativity sigma of gas), 5: product of atomic polarizabilities ion/gas, 6: Ion hydrogen bond acceptor x gas hydrogen bond donor, 7: Ion hydrogen bond donor x gas hydrogen bond acceptor. The first three combined properties correspond to attractive interaction if the product is negative (opposite charges of inter-component atoms). On the case of negative product will correspond the value of 1. If the product is 0 it will be associated the value of 0.5. Finally, if the product of charges is higher than 0 it will correspond the value of 0.

Each combined inter-component interaction property is normalized from 0-1 (seven properties). The inter-component interactions with a higher value of average, between the seven normalized combined properties, will correspond to more attractive interactions. Figures 2 and 3 highlights a first approach on the most important inter-component interactions within gas / cation and gas /anion on a reference IL + gas system. [6, 35, 36] According this first approach the most attractive cation/gas interactions correspond to core/adjacent imidazolium ring interacting with oxygen atoms of CO_2. The most attractive anion/gas contacts consist on carbon of CO_2 interacting with fluorine, oxygen, sulphur and nitrogen of the anion. Additionally, the interaction between oxygen atoms of CO_2 and carbon atoms of the anion are considered attractive.

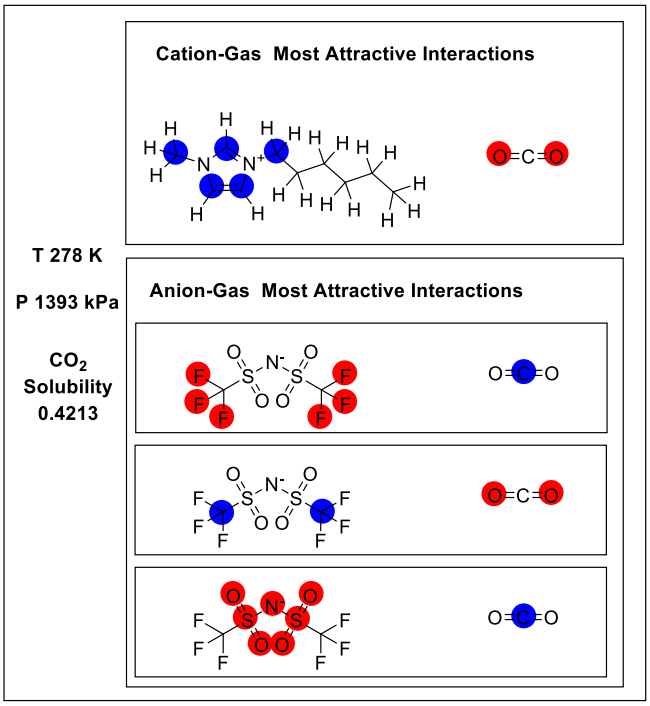


Figure 2. The most attractive interactions regarding Gas-Cation and Gas-Anion in a test set 1 data point system ([HMIM$^+$].[NTf$_2^-$] + CO_2). Red represent more electronegative/negatively charged atoms. Blue represent less electronegative positively charged atoms

It is worth to mention that the most repulsive interactions cation/gas are between core/adjacent imidazolium ring carbons and the carbon of CO_2. The anion/gas most repulsive contacts consist on interactions between oxygen atoms of CO_2 and fluorine atoms of the anion.

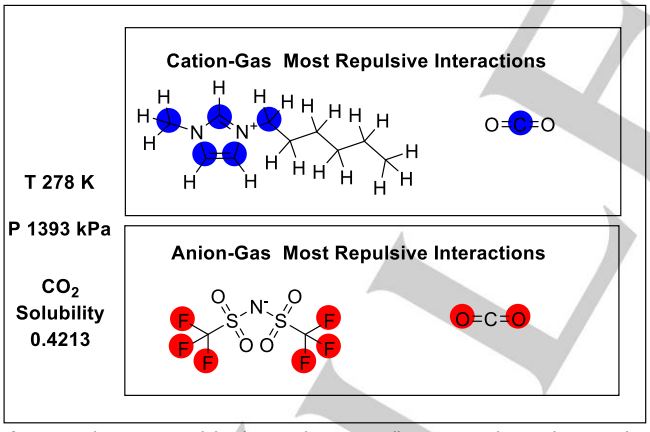


Figure 3. The most repulsive interactions regarding Gas-Cation and Gas-Anion in a test set 1 data point system ([HMIM$^+$].[NTf$_2^-$] + CO_2). Red comprise more electronegative/negatively charged atoms. Blue represent less electronegative positively charged atoms

It is important to highlight that the solubility of gases in ILs is not simply an occupation of free space within an ionic liquid, [5,37,38] otherwise different gases would present similar solubilities, which is not the case. Carbon dioxide and sulphur dioxide usually are highly soluble in ILs, differently O_2, N_2 and CH_4 are not. This fact strengthens the hypothesis that gas / IL interactions should provide an additional stabilization in order to rearrange the ionic liquid structure.[38] This chemoinformatic approach permits a fast, valuable estimation of gas solubility in IL's and, based on the new codification method, provide an initial approach on the most relevant inter-component atomic interactions ion/gas responsible for a certain solubility profile.

The chemoinformatic approach, provided by MOLMAP encoding method, highlights inter-component atomic gas/ion interactions. Differently, and in a complementary form, MD is an alternative way to investigate intra- and intermolecular interactions in a substance/mixture. By this method, atom-atom potential functions are used to generate molecular trajectories, which can then be analyzed in terms of attractive or repulsive forces between different groups of atoms in the molecules. Thus, in this work, the interaction profiles obtained from the MOLMAP results were compared with indications from molecular dynamics (MD) simulations performed at a temperature of 278 K and a pressure of 1393 kPa. These simulations were made using the CL&P and EPM2 models, [39-41] for which good reliability describing interactions between ILs and CO_2 molecules, was recently found against Nuclear Magnetic Resonance data. [42] The obtained results were analysed by the computation of the spatial distribution function (SDF) for the gas around the cations and anions, radial distribution functions (RDF), and studying molecular aggregation patterns.

It should be pointed out that, differently from MOLMAP, which gives information about the interactions between atoms of different components (cation/gas and anion/gas), MD simulations provide information regarding the global contacts between groups of atoms. Thus, it is difficult, for example, to separate the interactions of the carbon and fluorine atoms of the CF_3 groups of NTf_2^- and the atoms of the CO_2 molecule. Instead, the observed results provide a global contact view between all atoms as a group.

Figure 4a shows the arrangement of the CO_2 molecules around the cation charge centre (CC; imidazolium ring). This picture suggests that the CO_2 molecules are likely to be located above and below the imidazolium ring, or in front of the ring protons, similarly to MOLMAP encoding method in Figure 2. In addition, the surfaces are mainly located near the nitrogen atom connected to the methyl group. This is confirmed in the RDF obtained for the interactions between the nitrogen atoms of the imidazolium ring and the centre of mass of the CO_2 molecules. As can be observed in Figure 5a, the contact peak is more intense between the nitrogen atom connected to the methyl group, than with that attached to the alkyl chain. This is, most likely, the result of steric hindrance by the long chain.

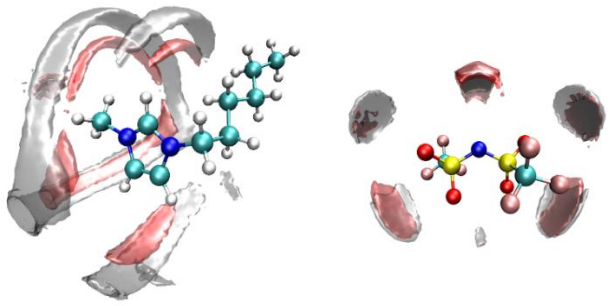


Figure 4. Spatial distribution functions for the distribution of the CO_2 molecules around the (a) cation imidazolium ring and (b) the anion centre of mass, from the molecular dynamics simulation results. The red and grey isosurfaces correspond to the distribution of the CO_2 oxygen and carbon atoms, respectively. The same cutoff was used to obtain the surfaces.

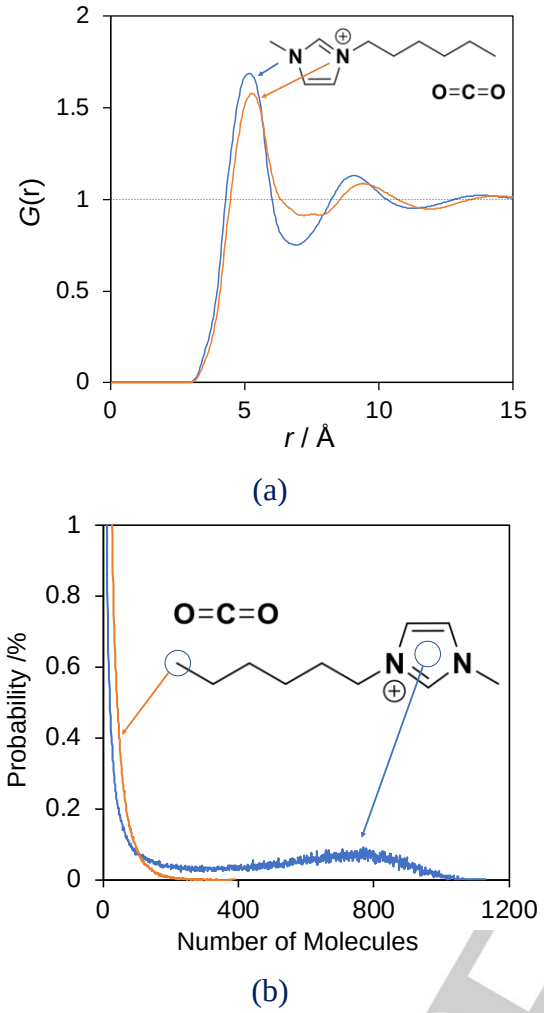


Figure 5. (a) Radial distribution function between the two nitrogen atoms of the IL cations and the centre of mass of the carbon dioxide molecules; (b) Probability of finding networks composed by cations (CAT) and CO_2 molecules, establishing alternating chains of the type ···CAT–CO_2–CAT– CO_2···, as a function of the number of elements in the aggregate. This analysis was performed assuming two contact centres in the cation: (*i*) the centre of charge – blue curve −and (*ii*) the terminal atom of the alkyl chain – orange curve.

Figure 4a also shows that the oxygen atoms of the CO_2 are closer to the CC of the cation than the carbon atom. This result is not unexpected since the global positive charge of $HMIM^+$ will preferentially interact with the negatively charged oxygen atoms of the gas molecules (note that, according to the EPM2 model, the carbon and oxygen atoms have atomic point charges of +0.6512 *e* and –0.3256 *e*, respectively). As a result, the CO_2 molecules approach the cation with their axis perpendicular to the CC, to maximize and minimize the interaction between the oxygen and carbon atoms, respectively. This is compatible with the conclusion obtained from the MOLMAP encoding approach in Figure 3, which suggests a repulsion between the carbon atoms of CO_2 and the carbon atoms of the cation aromatic ring. Another important feature highlighted by the MD results of Figure 4a, is the existence of interactions between the oxygen atoms of CO_2 and the imidazolium hydrogen atoms, suggesting the formation of hydrogen bonds between these atoms. The MOLMAP encoding approach indicates that such interactions are attractive, however not ranked among the most relevant attractive interactions.

The results also indicate a non-negligible interaction between the terminal carbon atom of the alkyl chain of the cation and CO_2. However, this interaction is less significant relative to that observed with the imidazolium group. The aggregation (contacts) between CO_2 and these two cation parts were investigated. For this analysis, it was assumed that a CO_2 is in contact with the cation CC or with the terminal carbon atom of the alkyl chain if their distance was smaller than 6.5 Å and 6.2 Å, respectively (criteria evaluated as previously described). [43] The obtained probability distribution as a function of the aggregate size is shown in Figure 5b. This image reveals that, while the CO_2 and the terminal alkyl group (CT) can form networks with up to 400 elements in alternating positions ···CT–CO_2–CT– CO_2···, if the charge centre is considered, equivalent networks with more than 1000 units can be observed. This suggests, therefore, that a significant interaction between CO_2 and the alkyl chain is also present. However, as suggested by the MOLMAP data, the stronger electrostatic interaction between the CO_2 molecules and the imidazolium group leads to a prevalence of this type of interaction.

Figure 4b gives the SDF plots for the distribution of the CO_2 molecules around the anions. From this image it can be concluded that: (*i*) the isosurfaces of the CO_2 carbon atoms are marginally closer to the anion than those found for the oxygens; (*ii*) the gas molecules tend to be located near the oxygen and nitrogen atoms of NTf_2^-, leading to the formation of five main interaction sites; (*iii*) no significant interactions between CO_2 and the fluorinated group are noticed. The first two observations are in line with the conclusions found from the MOLMAP encoding method discussed above in Figure 2, i.e., the main NTf_2^- – CO_2 interactions occur between the anion sulfoxide group. Also, as opposite interactions between CO_2 and the CF_3 groups can occur, negligible contacts between these groups are noticed.

The slight difference observed between the position of the isosurfaces obtained for the carbon and oxygen atoms of CO_2 around the anion is not unexpected. Due to the negative nature of the anion, the NTf_2^- - CO_2 interactions will necessarily involve the positively charged carbon atom of CO_2. As a result, unlike in the case of the cation, the CO_2 molecules face the electronegative atoms of the anion (oxygens and nitrogen) with the positively charged carbon atom. However, due to the existence of two opposing bonded (negative) oxygen atoms of CO_2, the gas molecules will approach the anion from the side, leading to similar isosurfaces when the oxygen and carbon atoms are considered.

The results obtained by the MD approach indicate that the MOLMAP encoding method and the average of the seven combined inter-component atomic interactions are an initial valuable tool to reveal attractive and repulsive interactions.

Figure 6, highlights the effect of the alkyl chain length of n-alkyl-methylimidazolium.tricyanomethanide on CO_2 solubility.

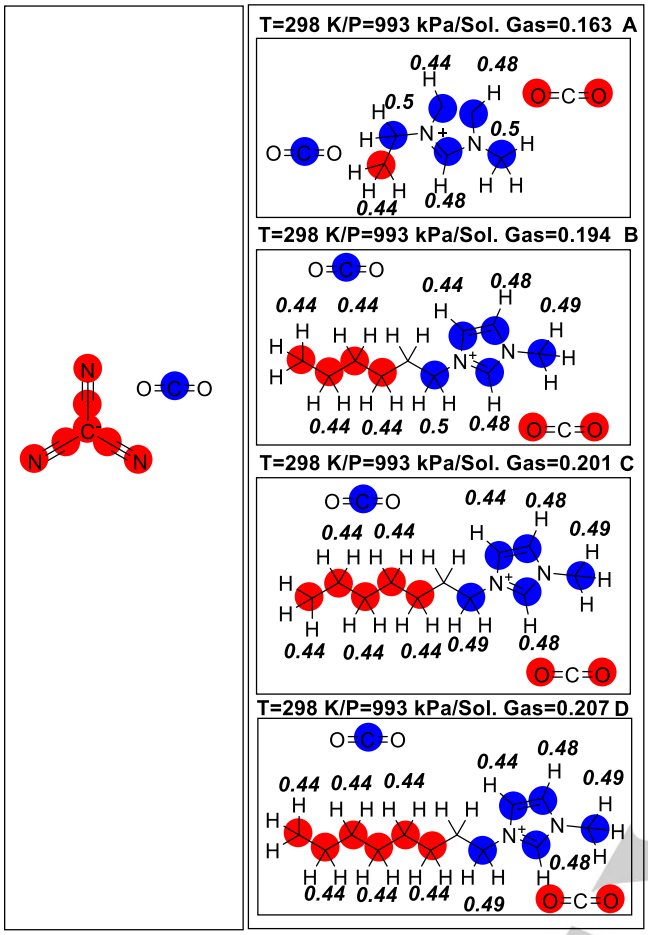


Figure 6. Effect of the size of n-alkyl chain of n-alkyl-MIM.tricyanomethanide on the CO_2 solubility. Visual identification of the most attractive interactions of cation/CO_2 at similar conditions of temperature and pressure. The values in italic represent the binary inter-component (cation-gas) atomic interactions with higher attractive index. Red comprise more electronegative/negatively charged atoms. Blue represent less electronegative positively charged atoms

It's possibe to observe that for similar temperature and pressure conditions the solubity of CO_2 (molar fraction) increases with the increment of the size of the alkyl chain. The observed trend may be due to the higher number of CH_2 groups, with a high attractive index, as the size of the alkyl chain increases (Figure 6), which means that when the size of the cation increases (extended alkyl chain), in combination with the increase of cation/gas attractive interaction sites, the mol number of CO_2 molecules per mol of IL increases. This observation is in line with the previously reported effect of alkyl chain length on CO_2 solubility. [5] This method highlights the reason for high difference of solubilities when CO_2 is compared with O_2 (identical ionic liquid TP conditions - Figure 7). Differently from the case of CO_2, the charges of O_2 are 0, contributing to low average value between the seven combined inter-component interactions (indexes of attraction ion gas) indicated in Figure 7.

The MOLMAP encoding method has been additionally tested to verify the differences of the most attractive interactions among ILs and CO_2 (constant cation and different anions Figure 8, A-D).

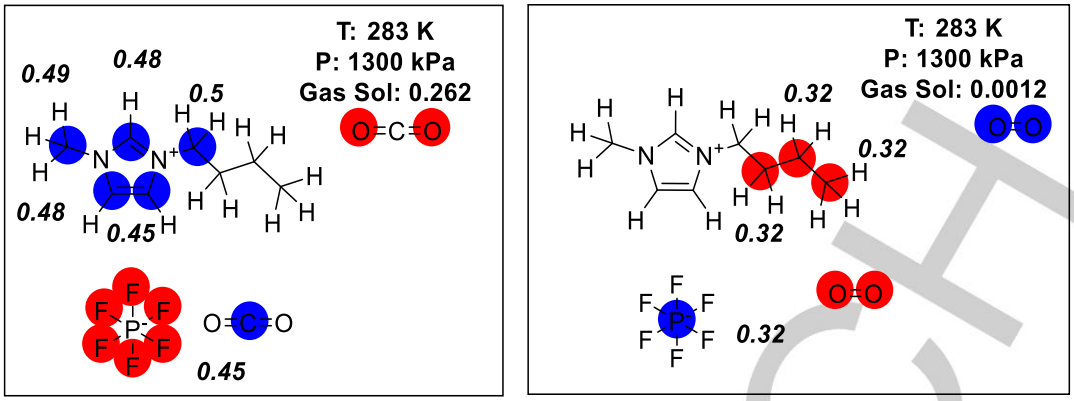


Figure 7. Visual identification of the most attractive interactions of BMIM.PF_6 dissolving CO_2 and O_2 at Identical condition TP. The values in italic represent the binary inter-component ion-gas atomic interactions with higher attractive index. Red comprise more electronegative/negatively charged atoms. Blue represent less electronegative positively charged atoms

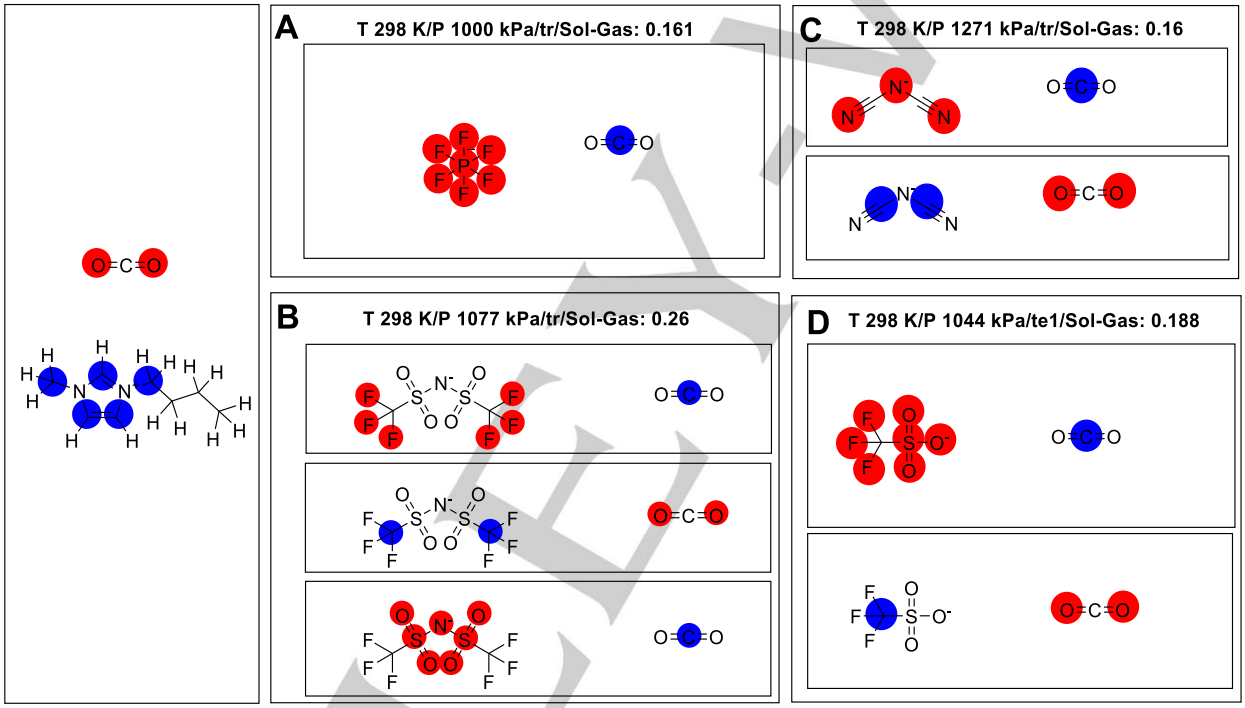


Figure 8. The effect of the anion on the different most attractive interactions anion/CO_2. PF_6 (A), NTf_2 (B), DCA (C), Triflate (D). Red comprise more electronegative/negatively charged atoms. Blue represent less electronegative positively charged atoms

A plausible explanation for the different solubility profiles observed in figure 8 (molar fractions) stems from the different sizes of the anions and the associated different number of atoms able to carry out attractive interactions. This approach explains the following order regarding the influence of anion on gas solubilities: NTf_2 > Triflate > PF_6 > DCA.

Conclusion

A new encoding method based on MOLMAP technology has been set. The method comprises, for the first time, the codification of combined atomic inter-component, cation/gas and anion/gas, interactions. The combination of the codification method and the Random Forest algorithm leads to robust models regarding predictive ability for new systems IL+Gas not present in the training set. The use of molecular dynamics validates the application of combined atomic inter-component cation/gas and anion/gas properties, used to encode MOLMAPs, as an initial valuable estimate of the most attractive/repulsive inter-component interactions. This method explains visually the difference between the solubilities of CO_2 and O_2 in identical IL at equal conditions TP. Finally, the effect of cation and anion are visually highlighted by this method in order to explain CO_2 solubility profiles. The increment of the ion size in combination with the higher number of atoms able to interact in an attractive form with the gas explain the obtained results.

This approach should find application on IL-based chemical system's design for gas separation, (namely O_2 purification in medical applications), greenhouse gas capture, reaction optimization, refrigerant gas absorption and antisolvent-based systems from both predictive and interpretation perspectives.

Methods

Dataset's construction and curation

Experimental Details. The extraction of solubility results of binary systems (IL + gas), from the ILThermo database, [44] has been carried out using the pyILT2 package. [45]

A selection of 6952 system's datapoints has been obtained after exclusion of non-VLE, repetitive and non-valid structures. These system's components (Cation, Anion and gas) have been standardized, using the standardizer [46] application from Chemaxon, (mesomerise, add implicit hydrogens and clean 3D structure).

This dataset was divided into a training set of 5615 system's datapoints, test set 1 with 702 datapoints and test set 2 containing 635 datapoints. The two independent test sets comprise datapoints where the combination of cation, anion and gas is not present in the training set combinations.

Chemical System's descriptors: MOLMAPs

The full standardized objects have been submitted to the Chemaxon cxcalc application [46] in order to calculate the profile of the atomic properties for each atom of the 6952 system's datapoints. These properties are obtained using the program ACPExtractMarvin application written in our labs. All the combinations of atoms between cation and gas / anion and gas, with the respective properties, have been gathered using a program built in JAVA language.

The selected inter-component combined properties consist on 1: atomic pi charge of ion x atomic pi charge of gas, 2: atomic sigma charge of ion x atomic sigma charge of gas, 3: atomic total charge of ion x atomic total charge of gas, 4: abs(atomic orbital electronegativity sigma of ion - atomic orbital electronegativity sigma of gas), 5: product of atomic polarizabilities ion/gas, 6: Ion hydrogen bond acceptor x gas hydrogen bond donor, 7: Ion hydrogen bond donor x gas hydrogen bond acceptor. The three first combined properties (charges) correspond to attractive inter-component interactions, if the product of the ion/gas charges is a negative value (associated the value of 1). If the product is 0, it corresponds to a value of 0.5. Finally, if the product is a positive value (a repulsive interaction of charges with identical sign), it is associated the value of 0.

The next step comprises the (0-1) normalization of each combined inter-component property, based on corresponding maximum and minimum present in the training set.

The next step comprises the random selection of 29999 atoms from the training set and training of a 30x30 *Kohonen* neural network. [47] This trained network is used to obtain patterns of activation comprising the inter-component anion / gas and cation / gas for the training set and both independent test sets (example of Figure 9). This is a generical form of grouping inter-component interactions based on the environment of a given atom, which may find resemblances between different elements or mark differences between identical elements, considering different chemical environments.

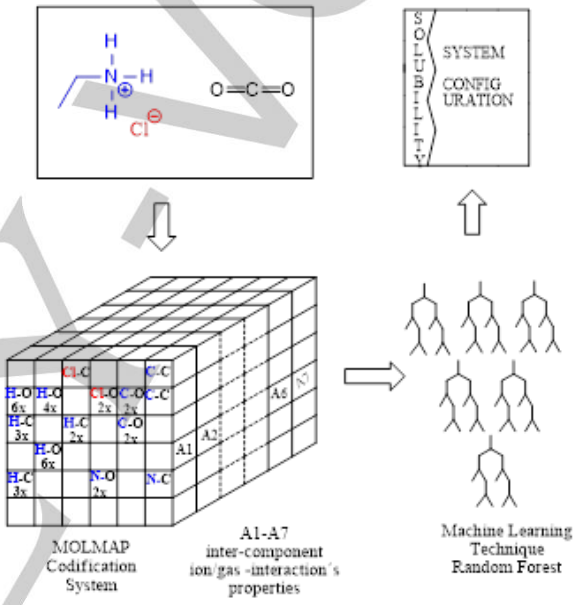


Figure 9. Chemoinformatic methodology, based on MOLMAP of intercomponent atomic interactions-encoding system and Random Forest machine-learning technology.

The patterns of activation are obtained first as an address for each inter-component (ion/gas) pair of atoms. Afterwards these addresses are converted into numerical descriptors using the program cutmapz3N, written in our labs. Each inter-component pair of atoms of a given system activates a winning neuron, a position of the *Kohonen* 30x30 matrix (MOLMAP), with a value 1. The neighbour neurons may also account a selected value, usually below the winning neuron value. Afterwards each position of the MOLMAP accounts for the number of times that the different inter-component pair of atoms of a system's datapoint activate it, or even, additionally, accounting the number of times that is considered as neighbour neuron for a given activation. The 30x30 matrix, representing the pattern of activation of each system's datapoint, is converted into a vector by concatenation of the lines of the matrix.

Machine Learning Technique:

The next step comprises the training of Random Forest (RF) models, with the training set. The RF algorithm is a set of predefined number of trees, each tree consisting on the partition of the training set objects from parent nodes into child nodes. The child nodes contain more resemblant datapoints, regarding the evaluated property (gas solubility), when compared with the respective parent node. Each partition is obtained by a logical rule based on a selected descriptor from a set of randomly selected *mtry* descriptors. Each tree is built with a randomly selected part of the training set. The remaining part is used to obtain the out of bag OOB predictions of that tree.The OOB predictions comprise the average value of solubility for a given system datapoint from trainning set (at certain temperature and pressure) for the trees where that generical datapoint was set aside on model construction. Finally, when the complete set of trees is obtained, each system datapoint submitted to the model will attain a value for prediction, the average value for all the trees, either a generical point from training or test set (Figure 9).

Random Forest (RF) [27, 28] is a machine learning technique, that permits multiple combinations of a given system's feature on a set of preselected number of trees. Due to that concrete characteristic, it has been applied in this work in combination with the MOLMAP codification system.

The next step is RF model tuning, where the MOLMAP profiles of activation per item (inter-component pair of atoms), winning neuron - first level immediate neighbourhood - second level - third level, are: 1-0-0-0, 1-0.5-0-0, 1-0.66-0.33-0 and 1-0.75-0.5-0.25, are evaluated. The *mtry* value is another optimization parameter.

A possibility offered by the encoding system, based on inter-component interactions ion/gas, consists on identifying for each mixture IL-gas which interactions are more attractive and more repulsive. The higher the average value among the seven combined normalized inter-component properties, the higher the attractive character of a given interaction.

Molecular dynamics validation

All simulations were performed using GROMACS 2019.4.[48] The initial system configuration was obtained by a random distribution of 1000 ion pairs of [HMIM$^+$][NTf2$^-$] and 728 molecules of CO_2, inside a cubic simulation box with a density d = 0.2 g.cm-3. This composition corresponds to a mole fraction of CO_2 of 0.42, which corresponds to the saturation limit of this gas in [HMIM$^+$][NTf$_2^-$].[49] The van der Waals and Coulomb interactions were computed assuming a distance of 1.6 nm. In the case of the Coulomb interactions, the particle-mesh Ewald technique was used to account for the electrostatic interactions beyond the cutoff. The system temperature (278.0 K) and pressure (1393 KPa) were kept constant by means of a Nose-Hover thermostat (relaxation time constant of 5 ps), and a Parrinello-Rahman barostat (relaxation time of 20 ps; compressibility of 4.5 × 10−5), respectively. The initial configuration was equilibrated by performing several 5 ns simulation runs until a constant density was observed. Finally, a production stage of 40 ns was made, recording the system configuration each 100 ps. For all simulation stages, a timestep of 2 fs was used. The ionic liquid force field was retrieved from the parametrization previously reported, [39,40] while the EPM2 model was selected to model the CO_2 molecules.[41] The input files for the simulations were prepared using Packmol [50] and DLPGEN 3.0. [51] The trajectory analysis was made using the software package AGGREGATES.[52]

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).

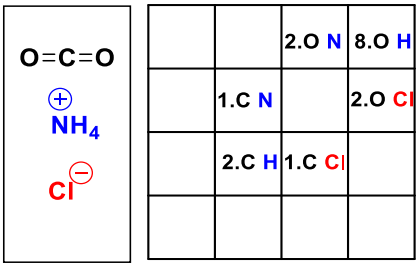
Prof. Dr. Manuel Nunes da Ponte and Prof. Dr. João Aires-de-Sousa are acknowledged for fruitful discussions.

Keywords: Ionic Liquids • Chemoinformatics • molecular dynamics • Gas • Solubility

[1] W. Beckner, C. Ashraf, J. Lee, D. A. C. Beck, J. Pfaendtner, *J. Phys. Chem. B*, **2020**, 124, *38*, 8347-8357.
[2] V. Venkatraman, S. Evjen, K. C. Lethesh, J. J. Raj, H. K. Knuutila, A. Fiksdahl, *Sustain. Energy Fuels*, **2019**, *3*, 2798-2808.
[3] G. Carrera, J. Aires-de-Sousa, *Green Chem*. **2005**, *7*, 20-27.
[4] A. Varnek, N. Kireeva, *J. Chem. Inf. Model*., **2007**, *47*, 1111-1122.
[5] Z. Lei, C. Dai, B. Chen, *Chem. Rev.* **2014**, *114*, 1289-1326.
[6] M. B. Shiflett, E. J. Maginn, *AIChE J.* **2017**, *63*, 4722-4737.
[7] K. Klimenko, J. M. Inês, J. Esperança, L. P. N. Rebelo, J. Aires-de-Sousa, G. V. S. M. Carrera *Mol. Inform.* **2020**. *39*, 2000001.
[8] K. Paduszynski, U. Domanska, *J. Chem. Inf. Model.*, **2014**, *54*, 1311-1324.
[9] G. V. S. M. Carrera, M. Nunes da Ponte, L. P. N. Rebelo, *ChemPhysChem*, **2019**, *20*, 1-8.
[10] D. Shang, X. Liu, L Bai, S. Zeng, Q. Xu, H. Gao, X. Zhang, *Curr. Opin. Green Sust. Chem.* **2017**, *5,* 74-81.
[11] J. P. Hallett, T. Welton, *Chem. Rev*. **2011**, *111*, 3508-3576.
[12] S. K. Shukla, S. G. Khokarale, T. Q. Bui, J-PT. Mikkola, *Front. Mater.*, **2019**, *6*, 42.
[13] S, Zeng, X. Zhang, L. Bai, X. Zhang, H. Wang, J. Wang, D. Bao, M. Li,, X. Liu, S. Zhang, *Chem. Rev*., **2017**, *117*, 9625-9673.
[14] B. R. Mellein, J. F. Brennecke, *J. Phys. Chem. B,* **2007**, *111*, 4837-4843.
[15] L. Dong, DX. Zheng, XH. Wu, *Ind. Eng. Chem. Res.,* **2012**, *51*, 4741-4747.
[16] F. Yusuf, T. Olayiwola, C. Afagwu, *Fluid Phase Equilib.*, **2021**, *531*, 112898.
[17] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.*, **2018**, *9 (10)*, 2639-2646.
[18] N. Goudarzi, D. Shahsavani, F. Emadi-Gandaghi, M. Arab Chamjangali, *Journal of Chromatography A*, **2014**, *1333*, 25-31.
[19] R. P. Sheridan, *J. Chem. Inf. Model.*, **2013**, *53*, 2837-2850.
[20] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, O. G. Kolumbin, N. N. Muratov, V. E. Kuz'min, *J. Chem. Inf. Model.*, **2009**, *49*, 2481-2488.
[21] M.Chen, K. Xiao, T. Zhao, Y. Zhou, Q. Zhang, J. Aires-de-Sousa, *J. Mol. Liq*., **2018**, *254*, 231-240.
[22] Q-Y. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.,* **2005**, *45*, 1775-1783.
[23] G. V. S. M. Carrera, S. Gupta, J. Aires-de-Sousa, *J. Comput. Aided Mol. Des*., **2009**, *23*, 419-429.
[24] D. A. R. S. Latino, J. Aires-de-Sousa, *Angewandte, Chem. Int. Ed*., **2006**, *45*, 2066-2069.
[25] B. Hemmateenejad, A. R. Mehdipour, P. L. A. Popelier, *Chem. Biol. Drug Des.*, **2008**, *72*, 551-563.
[26] S. Gupta, S. Matthew, P. M. Abreu, J. Aires-de-Sousa, *Bioorg. Med. Chem.*, **2006**, *14*, 1199-1206.
[27] L. Breiman, *Machine Learning*., **2001**, *45*, 5-32.
[28] https://cran.r-project.org/
[29] M. Jaschik, D. Piech, K. Warmuzinski, J. Jaschik, *Chem. Process Eng.*, **2017**, *38* (1), 19-30.
[30] N. A. Manan, C. Hardacre, J. Jacquemin, D. W. Rooney, T. G. A. Youngs, *J. Chem. Eng. Data* **2009**, *54*, 2005–2022.
[31] A. Yokozeki, M. B. Shiflett, *J. of Supercritical Fluids*, **2010**, *55*, 846-851.
[32] V. Venkatraman, B. K. Alsberg, *J. CO_2 Util.,* **2017,** *21,* 162-168.
[33] A. Baghbana, M. A. Ahmadi, B. H. Shahraki, *J. of Supercrit. Fluid.,* **2015**, *98,* 50-64.

[34] A. Shafiei, M. A. Ahmadi, S. H. Zaheri, A. Baghban, A. Amirfakhrian, R. Soleiman, *J. of Supercrit. Fluid*, **2014**, *95*, 525-534.
[35] K. N. Marsh, J. F. Brennecke, R. D. Chirico, M. Frenkel, A. Heintz, J. W. Magee, C. J. Peters, L. P. N. Rebelo, K. R. Seddon, *Pure Appl. Chem*., **2009**, *81*, 781-790.
[36] R. D. Chirico, V. Diky, J. W. Magee, M. Frenkel, K. N. Marsh, *Pure Appl. Chem.*, **2009**, *81*, 791-828.
[37] A. A. Oliferenko, P. V. Oliferenko, K. R. Seddon, J. S. Torrecilla, *Phys. Chem. Chem. Phys.,* **2011**, *13*, 17262-17272.
[38] S. P. Kelley, L. A. Flores, M. S. Shannon, J. E. Bara, R. D. Rogers, *Chem. Eur. J.*, **2017**, *23*, 14332-14337.
[39] A. S. L. Gouveia, C. E. S. Bernardes, L. C. Tome, E. I. Lozinskaya, Y. S. Vygodskii, A. S. Shaplov, J. N. Canongia Lopes, I. M. Marrucho, *Phys. Chem. Chem. Phys.*, **2017**, *19*, 29617-29624.
[40] J. N. Canongia Lopes, J. Deschamps, A. A. H. Padua, *J. Phys. Chem. B*, **2004**, *108*, 2038-2047.
[41] J. G. Harris, K. H. Yung; *J. Phys. Chem.,* **1995**, *99*, 12021-12024.
[42] M. Zanatta, M. Lopes, E. J. Cabrita, C. E. S. Bernardes, M. C. Corvo, J. CO_2 Util., **2020**, *41*, 1012252.
[43] K. Shimizu, C. E. S. Bernardes, J. N. Canongia Lopes, *J. Phys. Chem. B,* **2014**, *118*, 567-576.
[44] https://ilthermo.boulder.nist.gov/
[45] 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.
[46] https://chemaxon.com/
[47] T. Kohonen, *Biological Cybernetics*, **1982**, *4*, 59-69.
[48] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX*, **2015**, *1-2*, 19-25.
[49] S. Raeissi, L. Florusse, C. J. Peters, *J. Supercrit. Fluids,* **2010**, *55*, 825-832.
[50] L. Martínez, R. Andrade, E. G. Birgin, J. M. Martínez; *J. Comput. Chem*., **2009**, *30*, 2157-2164.
[51] C. E. S. Bernardes, A. Joseph; *J. Phys. Chem. A*, **201**5, *119*, 3023-3034.
[52] C. E. S. Bernardes, *J. Comput. Chem*., **2017**, *38*, 753-765.

Entry for the Table of Contents



This work comprises a Chemoinformatic approach based on the codification of all the pairs atomic intercomponent interactions between cation/gas and anion/gas (Ionic Liquid + Gas) in a *Kohonen* neural network (MOLMAPs). This form of codification, validated by molecular dynamics, is a first valuable tool to highlight the most attractive and repulsive interactions responsible for a concrete solubility of a given gas in an ionic liquid.