Machine-Learning Approaches to Tune Descriptors and Predict the Viscosities of Ionic Liquids and Their Mixtures

Gonçalo V. S. M. Carrera* and Manuel Nunes da Ponte[a]

This work consists on a new chemoinformatic approach based on two complementary artificial intelligence concepts. Random Forest and Kohonen neural network are applied on this context. The former provides a relevance measure of the numerical descriptors encoding either an ionic liquid or its mixtures. The code of a given chemical system is weighted according that relevance measure. The Kohonen neural network is trained with a set of weighted chemical systems. The next step comprises the use of the trained neural network as platform to obtain a tuned profile of numerical descriptors representing a generical chemical system. The tuning mechanism involves the topology of a chemical system-encoding vector in the neural network. The last step comprises the use of the tuned chemical systems to build predictive models of viscosities. The MOLMAP encoding technology is applied to represent ionic liquid systems and its mixtures.

1. Introduction

A given problem on bio-chemical-physical sciences requires a specific approach for a straightforward resolution. The antagonism of a bulk physical property and a targeted biological activity illustrates that concept when chemoinformatic methodologies are involved.[1] Different approaches are available regarding the study of biological activities.[2] The evaluation of bulk physical properties usually involves the compound/system as a unit, and ionic liquids (ILs) and ionic liquid-based mixtures illustrate the concept.[3] One conventional form of resolution involves the selection of the most relevant descriptors and, a plethora of techniques are available.[4] A different option consists on modifying the descriptor space to adjust it to a given problem. Principal Component Analysis (PCA) reduces the dimensionality and the descriptors become orthogonal.[5] Self-Organizing Maps (SOMs) are valuable tools on descriptor-size reduction and on finding meaningful relationships among available information.[6] The requirement of unified models that combine information of pure substances and mixtures is a challenge and recently our group proposed a method involving Molecular maps of atom-level properties (MOLMAP) applied on that context.[7] MOLMAPs have been tested with success in multiple applications.[8–12] This approach offers the possibility of generating chemical system-based descriptors in the form atomic/component pattern of activation in a Kohonen neural network, based exclusively on atomic/component generical property's profile.[13] This form of codification allows different atoms/components mapped in similar region of a Kohonen neural network considering that the physico-chemical environment presents a good matching. The aim of this work is a step-by-step model construction centred on the viscosity of diverse systems, either ionic liquid (ILs) and IL-based mixtures. Some previous approaches from different groups illustrate the challenge of modelling the viscosity of ILs, from base experimental measurements to the build-up of predictive models.[14–19]

The approach of this work, in a first step, comprises the use of the original MOLMAPs for the numerical codification of IL-based systems of different number/nature of compounds, considering its component's proportions for base-model construction centred on Random Forest (RF), similarly as in our previous work.[7] The second step corresponds to a descriptor tuning procedure involving Random Forest and Kohonen neural network algorithms. The later tunes the original MOLMAP descriptors, of a given system, based on its non-supervised form of learning relationships involving the entire systems of the training set. This methodology is described in detail in the Experimental Section. The accuracy performance of these models is accessed in Results and Discussion. The collected database used in this work is presented as supporting information. This work explores the combined/novel possibilities of different machine-learning algorithms conceived to obtain a more correct approach on finding low viscous IL-based systems. Viscosity is a fundamental property of ILs decisive on its domains of applicability. The viscosity accessibility by predictive methods focus time and resources on the preparation of interesting examples for a given application.

2. Results and Discussion

This work comprises the evaluation of two different predictive models for classification of viscosity of ionic liquids and its mixtures (Check Experimental Section). Both models involve six classes (A–F) in ascending order of viscosity. These models are

[a] Dr. G. V. S. M. Carrera, Prof. M. Nunes da Ponte
Associate Laboratory for Green Chemistry-LAQV
Faculty of Sciences and Technology
Nova University
2829-516 Caparica (Portugal)
E-mail: goncalo.carrera@fct.unl.pt

based on Random Forest with 1000 trees and *mtry* = 75, set.seed(100) using R-3.5.3. The descriptor's dimensionality and character, representing numerically a generical chemical system, are MOLMAP 20×20 (400) atom + MOLMAP 12×12 (144) component + temperature (T) + pressure (P) in both cases, (400 + 144 + 2 = 546 descriptors). Both models are trained with the ionic liquid-based systems of the training set (6937 systems). The difference between the two models is that the base version is built with the original MOLMAPs, T and P. The novel version, the aim of this work, is based on tuned MOLMAPs, T and P by the combined use of Random Forest and a trained Kohonen neural network. Random Forest provides to a given descriptor position a certain ponderation factor, based on importance measures. The trained Kohonen Neural Network provides a match between a given chemical system-descriptor's profile and a winning neuron of the network set of weights (lowest Euclidian distance). The localization of a given object in the Kohonen network allows one to establish the first-level neighbourhood of the winning neuron. Each neuron of the first-level neighbourhood has a concrete Euclidian distance with a chemical system-descriptor's profile. The next step comprises a balanced average considering, the original set of MOLMAP + T + P descriptors of a chemical system, with the highest ponderation factor; the set of weights of the corresponding winning neuron, with the second highest ponderation factor; and finally, the set of weights of each neuron of the first level neighbourhood with a ponderation factor inversely proportional to the Euclidian distance. The train of a Kohonen neural network comprises the sequential submission of each chemical system of the training set to the network, each time a winning neuron is found, its weights and neighbourhood-neuron's weights are adjusted in order to become more similar to that concrete chemical system set of descriptors. After the training the Kohonen neural network set of weights, and its localization, is based on the complete training set. The form of tuning a given chemical system set of descriptors is based on the possibility that all training chemical systems used to train a Kohonen neural network contribute to that end.

Table 1 compares the base RF model with the RF tuned descriptor's model version. The modelling results and the Out of Bag (OOB) validation for 6937 systems shows that the RF tuned descriptor's model is more effective to obtain higher accuracies and percentages of correct and adjacent assignments. The opposite trend is followed by validation set 1, with 767 systems. Finally, the validation set 2 with 366 systems presents similar results with both methods. This evaluation doesn't permit to identify what is the more efficient model. It's here highlighted that our validation sets 1 and 2 comprise systems where at least one component, either ion or molecule, is different from the combinations existent in the training set. This is a rigorous form of evaluating the predictive ability of our models. Billard, Marcou, Ouadi and Varnek,[16] tested back-propagation neural networks and fragment molecular descriptors to encode variable cation and anion. The approach is innovative however, the results are modest. Other authors tested linear[20–25] and non-linear[20,24] methods in order to correlate and predict viscosities of single ILs at variable temperature[20–25] and also pressure[20,23,24] with acceptable predictive ability. The test set selection form is random.

A different approach consists on the use group contribution-based methods in order to predict viscosities of single ILs,[15,18,19] however the selection of groups for a given IL is time-consuming.

Mixtures involving ILs have been tested by different authors at variable temperature[26,27] and also pressure.[26] Good fit is obtained when the model is correlative[26] or the test set selection is random.[27]

Our approach, differently from those works, is based on fixed length matrixes and allows one to compare systems of different number/nature of components at variable temperature and pressure. The method is not limited to a given number of compounds in a chemical system (virtually infinite number of compounds), differently from the previous approaches. The build up of predictive model is automatic and the test set selection criteria assures that a given chemical system's combination of components (cation, anion and molecule) in the test sets is different when compared to the training set. This is a classification model that recognizes, in a straightforward form, low viscosity systems (Table 2).

A different parameter should be considered when an evaluation involving the base and tuned methods is carried out. It consists on the capacity to recognize low viscosity systems - class A objects. This form of evaluation allows one to find what is the more suitable model to identify low viscosity systems which opens new perspectives in order to find applications for ionic liquids.

Table 2 shows clearly, for all forms of validation, that the tuned-descriptor's model recognizes more accurately low viscosity systems (Check example of validation 1 - Contingency tables – Tables 3 and 4). This result shows that the RF tuned-descriptor's model is more able to carry out this recognition than the base RF model. Check, additionally, all the contingency tables in the supporting information.

Table 1. Comparative study between the base Random Forest (RF) model and equivalent tuned descriptor's RF model. Accuracy and % of Correct and Adjacent assignments.

Set	Base RF model		Tuned-descriptors RF model	
	Accuracy (% Correct)	% Correct and Adjacent	Accuracy (% Correct)	% Correct and Adjacent
Training (6937 systems)	91.11	99.57	99.99	100
Validation Out of Bag (6937)	70.69	99.11	79.31	99.19
Validation 1 (767)	65.19	92.18	58.80	90.48
Validation 2 (366)	62.02	87.16	61.20	89.89

Table 2. Comparative evaluation of base RF model and tuned model. Percentage of correct assignments, Class A and Not Class A.

Set	ase RF model		Tuned-descriptor's RF model	
	% Correct Class A	% Correct Not Class A	% Correct Class A	% Correct Not Class A
Training (6937 systems)	95.66	89.97	100	99.98
Validation Out of Bag (6937)	88.70	66.22	92.03	76.15
Validation 1 (767)	78.76	62.84	91.15	53.21
Validation 2 (366)	90.24	47.74	88.62	47.32

Table 3. Validation 1 Tuned descriptor's model contingency table.

		Experimental					
		A	**B**	**C**	**D**	**E**	**F**
Predicted	**A**	103	24	10	2	0	0
	B	7	78	41	14	1	0
	C	2	16	94	42	5	3
	D	1	8	36	134	44	21
	E	0	0	1	10	23	12
	F	0	0	0	5	11	19

Table 4. Validation 1 Base model contingency table.

		Experimental					
		A	**B**	**C**	**D**	**E**	**F**
Predicted	**A**	89	23	9	1	0	0
	B	22	74	25	13	2	0
	C	2	23	124	24	6	9
	D	0	6	23	163	43	10
	E	0	0	1	5	30	16
	F	0	0	0	1	3	20

Figure 1 presents diverse examples showing that the tuned descriptor's RF model correctly identifies class A viscosity and the base model fails on that task

The tuned-descriptor's RF model has been validated by three-fold cross-validation leading to an accuracy of 74.61%. Finally, the model has been validated by randomization of the classes in the training set and prediction for validation sets 1 and 2 (5 assays). The accuracy ranges from 16–25% and 16–28%, respectively. These results show that there's an intrinsic order between the tuned descriptors and the class of viscosity.

Figure 2 visually describes the effect of the chain length on 1-alkyl-3-methyl imidazolium cation. It's observable an increase of viscosity with the increment of the size of the chain. The formation of two segregated phases (coloumbic vs apolar) and the associated lower mobility between components explains the behaviour of viscosity.

Figure 3 highlights the effect of the blockage of the 2-position of 1-butyl-3-methylimidazolium by a methyl group. When the 2-methyl group is absent this position is moderately acidic contributing to a close interaction with the anion forming an ionic pair more easily and neutralization of charge, leading to lower viscosities. The 2-methyl group is much less acidic, and the ionic pair is unlikely to form.

Figure 4 comprises the effect of the anion on the viscosity of ILs and it's observable that small anions with high degree of delocalization of charge contribute to low viscosities. Contrarily centred charge, high dimension-long chain anions lead to high values of this property.

Figure 5 identifies another isomerism effect – the ramification. When the ramification increases the rotation degrees of freedom decrease, the alkyl chains become more rigid and a resultant lower mobility/higher viscosity is obtained.

Figure 6 shows the effect of the cation core on the viscosity. It's possible to observe that the imidazolium core leads to lower viscosities as the 2-position of the ring is moderately acidic forming more easily an ionic pair with the anion, reducing the coulombic effect on the mobility of the salts.

The complete characterization of the datasets is carried out in the supporting information. It is important to highlight, considering the included information, that the pure chemical systems include 1 cation and 1 anion with the sum of 2. This is the converted compound to component molar fraction. The value of the sum of components has been set to two and involve pure systems, mixtures of two ionic liquids and mixtures of an ionic liquid and a molecule.

This methodology is able to correctly identify the effect of different structural features on the viscosity classification of ionic liquid-based systems. The capacity of recognizing low viscosity systems is amplified with the use of this tuning procedure.

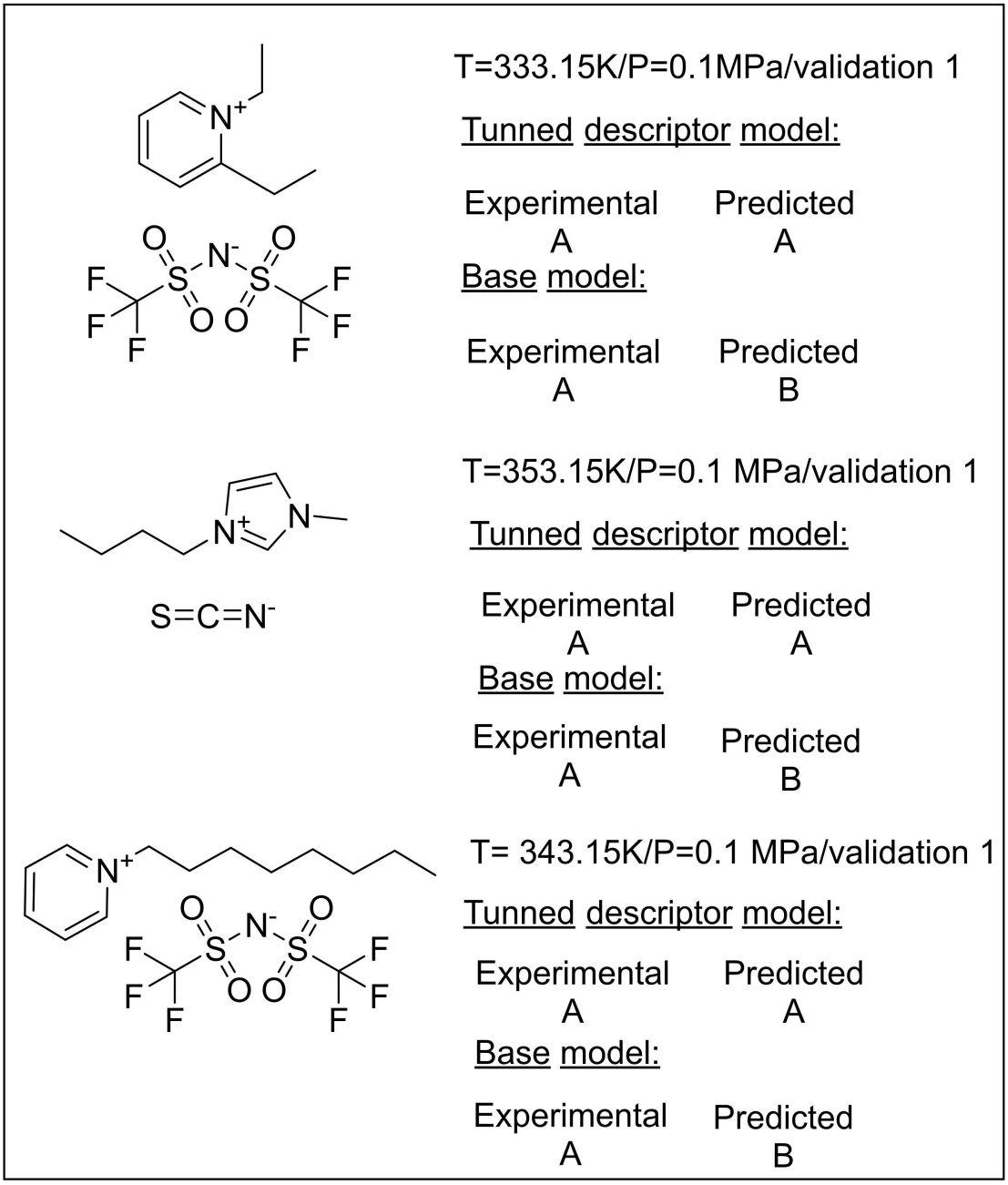

Figure 1. Examples of systems showing that the tuned-descriptors RF model more correctly distinguish true class A when compared to base RF model.

3. Conclusions

Two different methodologies have been applied in order to predict the viscosities of ionic liquids and its mixtures at different conditions of temperature and pressure. The MOLMAP descriptor technology, characterizing those chemical systems, has been applied on this context. A base Random Forest model has been trained with ionic liquids and its mixtures characterized by simple MOLMAP descriptors, temperature, and pressure. A different Random Forest model has been built with tuned MOLMAP + temperature + pressure descriptors. This tuning procedure is based on two different machine learning algorithms, a Random Forest supervised method and a non-supervised Kohonen neural network methodology. The tuned-descriptor method recognizes in a more straightforward form low viscosity systems than the base model. The tuned-descriptor method recognizes in a straightforward form the structural differences between comparable ionic liquids.

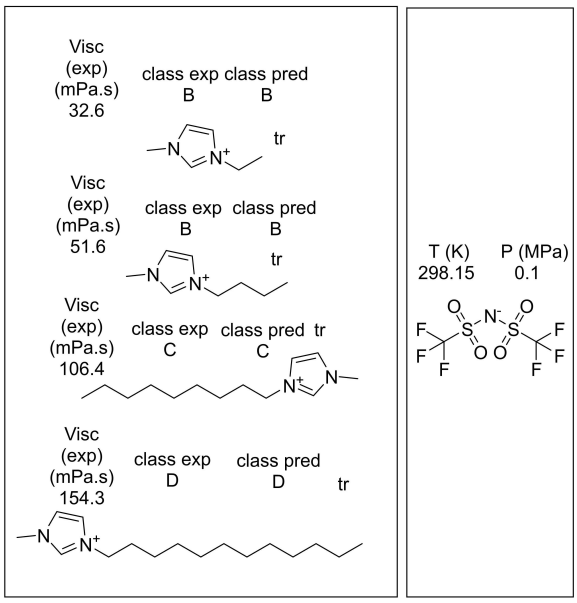

Figure 2. Effect of the chain length of the cation.

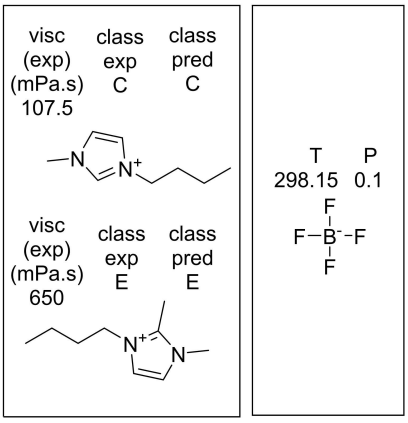

Figure 3. Effect of 2-methyl-group functionalization on the viscosity of 1-butyl-3-methyl imidazolium.

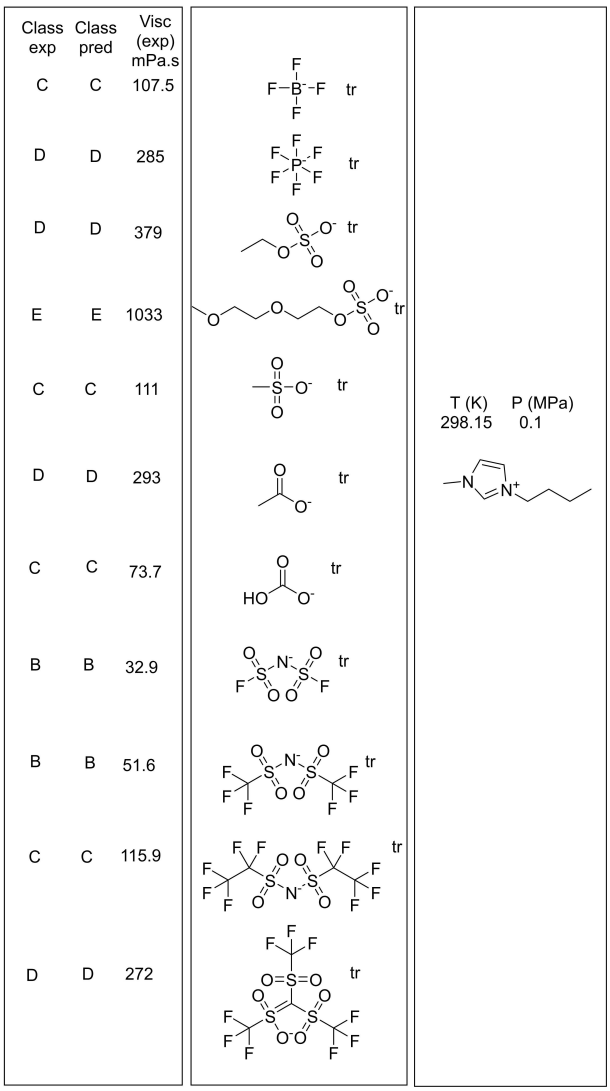

Figure 4. Effect of the anion on the viscosity of ionic liquids.

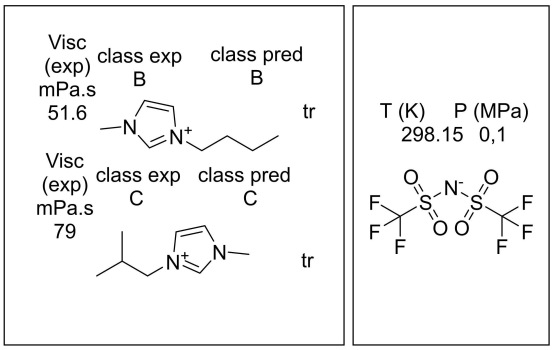

Figure 5. Isomerism 2 – Ramification effect.

Experimental Section

Collection of Data

A collection of 13798 chemical systems,[28] with the respective viscosity, has been obtained from NIST ILThermo database.[29] This collection includes pure ionic liquids, mixtures of two ionic liquids and mixtures of an ionic liquid and a molecule. A selection of 8070 chemical systems have been obtained in order to avoid errors and redundancies. This selection has been distributed into a 6937-chemical system's training set, and two validation sets of 767 and 366 chemical systems. Both validation sets comprise chemical systems where at least one component is not present in a generical combination of the training set's chemical systems.[30]

Standardization of Structures and Estimation of Atomic and Component Properties

Each component's chemical structure, either a cation, anion and molecule, encoded in smiles format, has been processed using the Chemaxon platform.[13] This is a two-step operation involving the a) Standardization of structures with standardizer and b), c) determination of physico-chemical properties of components and atoms respectively with cxcalc:

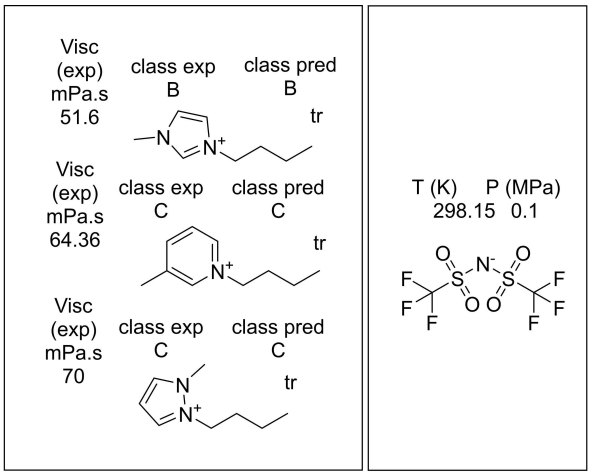

Figure 6. Cation core effect.

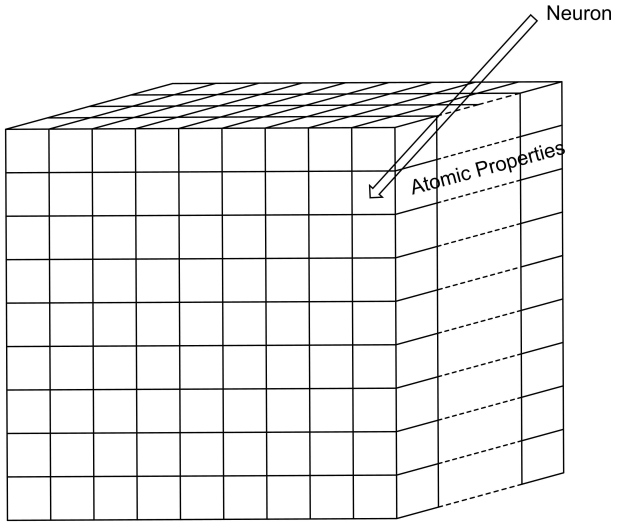

Figure 7. Generical atom-trained Kohonen neural network. Similar architecture for component-trained neural network. Each neuron is one cell in the surface of the network.

a) Standardization involves the following sequence of operations: 1) Mesomerize, 2) Add explicit H and 3) Clean 3D.

b) Surface area, polarizabilities, volume and weight, within a component

c) Total charge, pi charge, sigma charge, hydrogen bond donor, hydrogen bond acceptor, hindrance, atomic number from elements, polarizabilities, and orbital electronegativity sigma.

All the properties have been normalized 0–1.

Generation of MOLMAP Descriptors

A set of 5000 components has been chosen randomly from our collection of chemical systems in the training set: 2413 anions, 2307 cations and 280 molecules. This characterized dataset has been used to train component-type Kohonen networks according the component's normalized properties profile (Experimental Section B). The number of component's properties is identical to the number of levels, (weights) of each neuron's Kohonen neural network used to obtain MOLMAP system component's profile. A set of 10000 atoms has been selected randomly in order to train atom-type Kohonen network: 4500 atoms of cation, 4000 from anion and 1500 atoms from molecules. The number of weights, representing the deepness of each neuron of the Kohonen neural network is identical to the number of atomic properties representing each atom (Figure 7). The train consists on submitting objects of a certain type, either components or atoms, to a generical Kohonen neural network. Each object activates a neuron, the wining neuron. It's the neuron with the lowest Euclidian distance, comparing the object-property's profile with the set of weights of that neuron. The weights of the winning neuron are adjusted for a higher resemblance respective to the object-property's profile. The weights of neighbour neurons are adjusted according the distance to the winning neuron. All the objects of the training set are submitted sequentially a selected number of times.

The component-type networks comprise 12×12 (144 descriptors), 15×15 (225) and 18×18 (324) trained Kohonen networks. 20×20 (400 descriptors), 25×25 (625) and 30×30 (900) atom-type Kohonen neural networks have been built.

The trained Kohonen neural networks, for atoms and components are used to obtain MOLMAP numerical descriptors:

Each component, of a generical chemical system, (ions, molecules), from pure ionic liquids, mixtures of two ionic liquids, and mixtures of an ionic liquid with a molecule, has associated a value of component molar fraction. Each component of a chemical system is represented by its atoms. Its atoms are submitted to a generical-trained atomic Kohonen neural network.

Each atom activates a neuron with the value of 1 (the winning position in the Kohonen neural network). The neighbour neurons activate a lower value. The activation value of 0.5 has been tested for the first level neighbourhood.

Each chemical system's component has its pattern of atomic activation in the Kohonen neural network.

The next step comprises the product between the activation value of each position (neuron), in the Kohonen neural network (pattern of activation), and the component molar fraction.

This procedure is carried out for all the components of a generical chemical system.

The atomic molar-fraction weighted pattern of activation of all the components of a generical chemical system, either a pure ionic liquid, mixture of two ionic liquids or a mixture between an ionic liquid and a molecule, are summed, resulting in the MOLMAP of atoms of a generical chemical system (Figure 8).

Identical procedure is followed for MOLMAP of components. The difference consists on the use of components, instead of atoms, submitted to the trained Kohonen neural network of components.

The MOLMAP of either, atoms or components, is converted to vector from matrix, by sequential concatenation of the lines of the matrix

Different combinations of MOLMAPs of atoms and components have been tested for viscosity modelling in our previous work,[7] as a set of descriptors, in combination with temperature and pressure. The variables tested and optimized were the size of the Kohonen neural network, and inclusion of neighbourhood activation by a given item, either a component or an atom. The results of the previous work indicate that atomic 20×20 in combination with 12×12 component descriptors, corresponding to activation of 1–0.5 for the winning and neighbour neurons, respectively, are the most

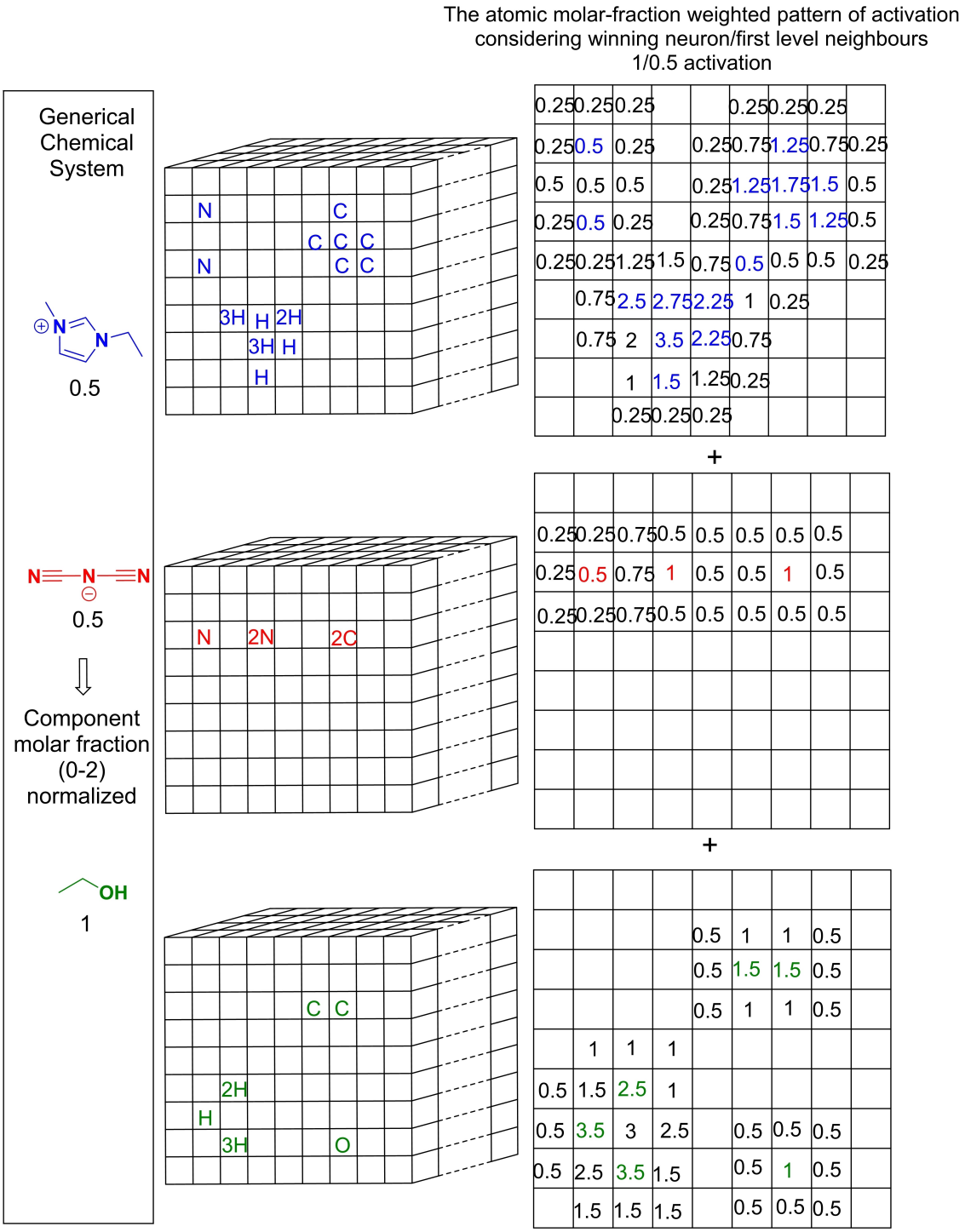

Figure 8. Schematic procedure to obtain MOLMAP of atoms of a generical chemical system (mixture of an ionic liquid and a molecule).

adequate descriptors for the next stage, the tuning of system-based descriptors.

Tuning System-Based Descriptors

An initial/base viscosity RF[31] model is built comprising 1000 trees and *mtry* = 75, optimized with out of bag (OOB) accuracy. The model is based on a pool of 20x20 (400) atomic, 12×12 (144) component, temperature and pressure descriptors, characterizing each system, corresponding to an activation of 1–0.5 for the winning and first-level neighbour neurons respectively, for each atom/component activation moiety. This is a classification model with six (A–F) classes: Class A: 0.28–20.59 mPa.s, Class B: 20.6–51.6 mPa.s, Class C: 51.7–122.9 mPa.s, Class D: 123–414 mPa.s, Class E: 415–1035.5 mPa.s and Class F: 1036–140000 mPa.s. The RF algorithm provides a scale of descriptor importance (each position of MOLMAPs of atoms and components, pressure and temperature). This scale is normalized from 0–2. The descriptors are normalized

from 0–1, and multiplied by the 0–2-corresponding normalized importance. All the sets are normalized and weighted according this procedure (Figure 9).

The tuning procedure's first step comprises the train of a 30×30-neuron Kohonen network with the normalized/weighted chemical systems of the training set. Afterwards we have access to the weights of the network, each level of weights corresponds to a given descriptor position and each neuron has as many levels/weights as number of descriptors. This trained network is the reference to tune the descriptors of all system's datasets (Figure 10):

i) Each system has its own wining neuron, considering this fact, it's compared the weights of that neuron and that corresponding system's descriptors in the form of Euclidean distance

ii) Identical procedure is followed for each of the eight neurons of the first level neighbourhood of that concrete system's winning neuron. Each corresponding neighbour neuron, set of weights, compares with a that generical chemical system's descriptors identically as in i).

iii) Each system of a given dataset will have associated, afterwards, the corresponding values of Euclidean distances for the winning neuron and first level neighbourhood.

The second step of this tuning procedure comprises:

a)→All the existent values of Euclidean distance were normalized from 0–1, considering that a system compared with itself corresponds to 0 and it's the minimum value for normalization.

b)→The normalized Euclidean distances have been converted to ponderation weights by subtraction of the Euclidian respective to the value of 1.

c)→The system descriptor's vector, wining neuron set of weights, and each first-level neighbour neuron set of weights is multiplied, respectively, by the reciprocal euclidian distance-based weights of:

- that system,

- winning neuron and,

- the eight neighbour neurons.

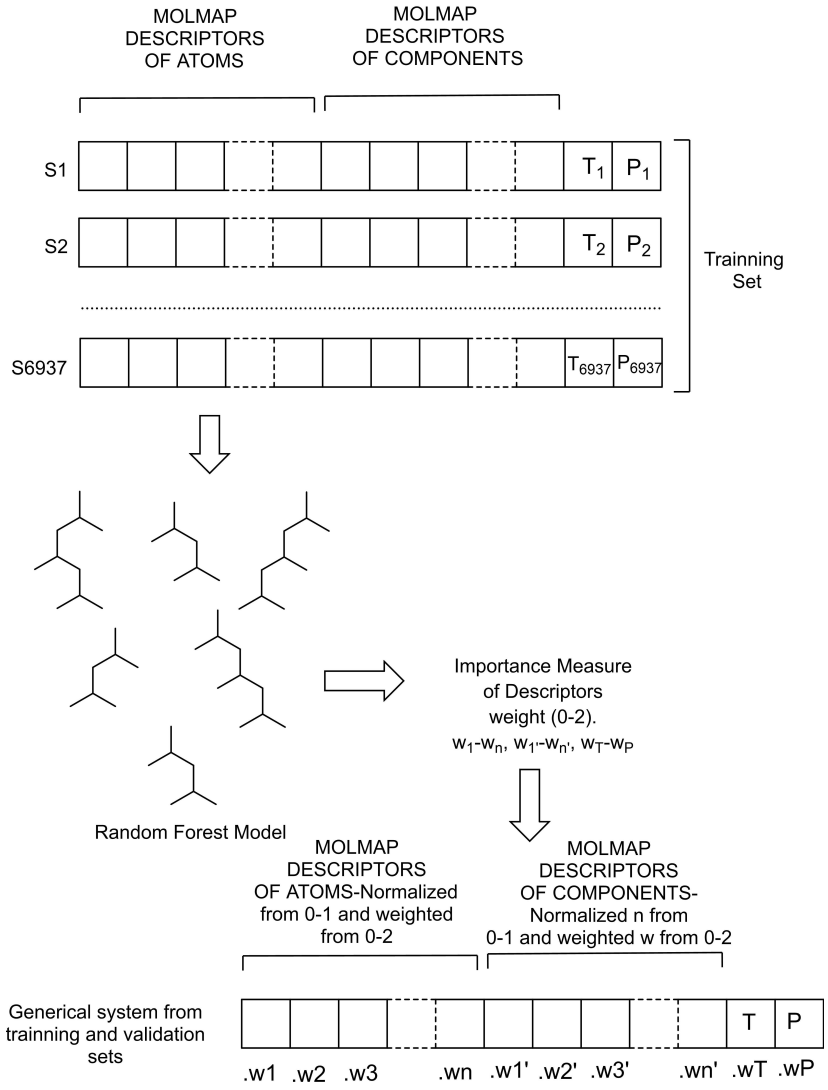

Figure 9. Procedure for 1–0 normalization and 0–2 weighting of a generical chemical system, from training and validation sets, based on importance measures from trained Random Forest.

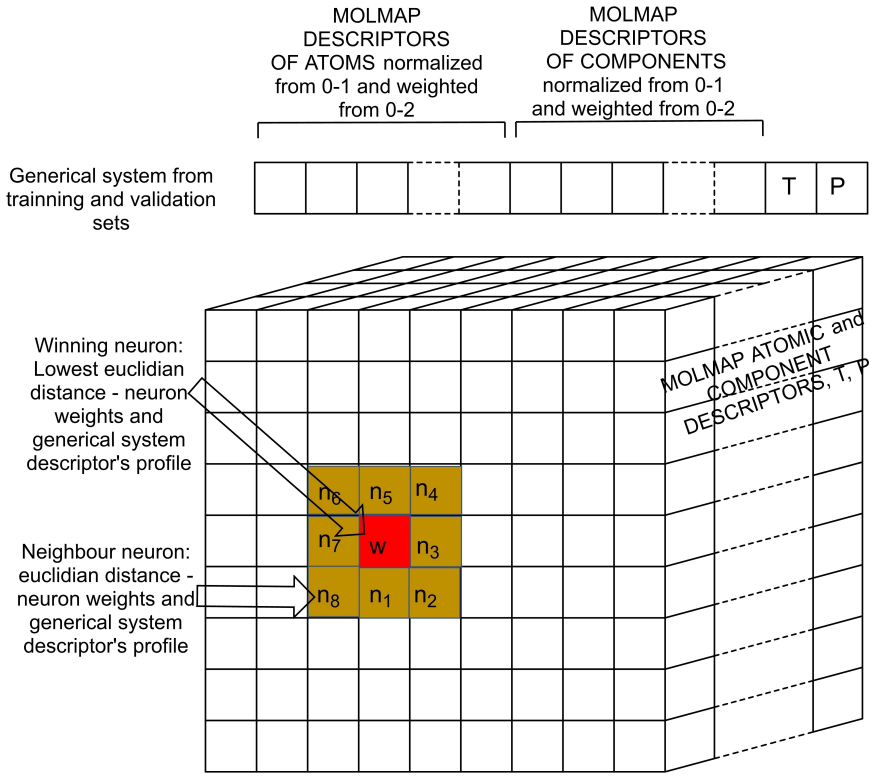

Figure 10. First step tuning procedure of a generical chemical system based on a trained Kohonen neural network.

d)→Each chemical system is represented by these ten weighted vectors, which are summed and divided by the sum of ponderation weights (of that given system, winning neuron and the eight first-level neighbour neurons). All the chemical systems are tuned using this procedure (Equation 1).

$$\text{Tuned descriptor's profile} = \frac{1.\text{system's descriptorprofile}+x.w+x1.n1+x2.n2+x3.n3+x4.n4+x5.n5+x6.n6+x7.n7+x8.n8)}{(1+x+x1+x2+x3+x4+x5+x6+x7+x8)}$$

w:winning neuron weights x: 1 - normalized Euclidian distance (w) winning neuron weights vs system's descriptor profile
n1: neighbour 1 weights x1: 1 - normalized Euclidian distance (n1) neighbour neuron weights vs system's descriptor profile
n2: neighbour 2 weights x2: 1 - normalized Euclidian distance (n2) neighbour neuron weights vs system's descriptor profile
....
n8: neighbour 8 weights x8: 1 - normalized Euclidian distance (n8) neighbour neuron weights vs system's descriptor profile

Equation1. Tuning procedure - second step

Finally, a new Random Forest model is built and validated with tuned training and validation datasets, respectively. The 6 classes are identical as in the Random Forest-base model.

Acknowledgements

Portuguese Foundation for Science and Technology Project: PTDC/EQU-EQU/30060/2017.

Prof. Dr. João Aires de Sousa for fruitful discussions.

This work was supported by the Associate Laboratory for Green Chemistry LAQV which is financed by national funds from FCT/MCTES (UIDB/50006/2020)

Conflict of Interest

The authors declare no conflict of interest.

Keywords: chemoinformatics · ionic liquids · Kohonen neural network · random forest · viscosity

[1] T. Engel, J. Gasteiger, *Applied Chemoinformatics*, Wiley-VCH, Weinheim, **2018**, pp 9–52.
[2] a) J. P. A. Martins, E. G. Barbosa, K. F. M. Pasqualoto, M. M. C. Ferreira, *J. Chem. Inf. Model.* **2009**, *49*, 1428–1436; b) X. Dong, J. O. Ebalunode, S. J. Cho, W. Zheng, *J. Chem. Inf. Model.* **2010**, *50*, 240–250.
[3] a) G. Carrera, J. Aires-de-Sousa, *Green Chem.* **2005**. *7*, 20–27; b) A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin, V. P. Solov'ev, *J. Chem. Inf. Model.* **2007**, *47*, 1111–1122.
[4] M. Shahlaei, *Chem. Rev.* **2013**, *113*, 8093–8103.
[5] T. Engel, J. Gasteiger, *Chemoinformatics*, Wiley-VCH, Weinheim, **2018**, pp 399–437.
[6] P. Schneider, A. T. Meller, G. Gabernet, A. L. Button, G. Posselt, S. Wessler, J. A. Hiss, G. Schneider, *Mol. Inf.* **2017**, *36*, 1600011.
[7] G. V. S. M. Carrera, M. Nunes da Ponte, L. P. N. Rebelo, *ChemPhysChem* **2019**, *20*, 2767–2773.
[8] S. Gupta, S. Matthew, P. M. Abreu, J. Aires-de-Sousa, *Bioorg. Med. Chem.* **2006**, *14*, 1199–1206.
[9] G. V. S. M. Carrera, S. Gupta, J. Aires-de-Sousa, *J. Comput.-Aided Mol. Des.* **2009**, *23*, 419–429.
[10] D. A. R. S. Latino, J. Aires-de-Sousa, *Angew. Chem. Int. Ed.* **2006**, *45*, 2066–2069; *Angew. Chem.* **2006**, *118*, 2120–2123.
[11] D. A. R. S. Latino, Q.-Y. Zhang, J. Aires-de-Sousa, *Bioinformatics* **2008**, *24*, 2236–2244.
[12] Q.-Y. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2005**, *45*, 1775–1783.
[13] https://chemaxon.com/.
[14] G. Yu, D. Zhao, L. Wen, S. Yang, X. Chen, *AIChE J.* **2012**, *58*, 2885–2899.
[15] K. Paduszyński, U. Domańska, *J. Chem. Inf. Model.* **2014**, *54*, 1311–1324.

[16] I. Billard, G. Marcou, A. Ouadi, A. Varnek, *J. Phys. Chem. B* **2011**, *115*, 93–98.
[17] S. Martin, H. D. Pratt, III, T. M. Anderson, *Mol. Inf.* **2017**, *36*, 1600125.
[18] R. L. Gardas, J. A. P. Coutinho, *Fluid Phase Equilib.* **2008**, *266*, 195–201.
[19] R. L. Gardas, J. A. P. Coutinho, *AIChE J.* **2009**, *55*,1274–1290.
[20] Y. Zhao, Y. Huang, X. Zhang, S. Zhang, *Phys. Chem. Chem. Phys.* **2015**, *17*, 3761–3767.
[21] Z. K. Koi, W. Z. N. Yahya, R. A. A. Talip, K. A. Kurnia, *New J. Chem.* **2019**, *43*, 16207–16217.
[22] S. A. Mirkhani, F. Gharagheizi, *Ind. Eng. Chem. Res.* **2012**, *51*, 2470–2477.
[23] F. Yan, W. He, Q. Jia, Q. Wang, S. Xia, P. Ma, *Chem. Eng. Sci.* **2018**, *184*, 134–140.
[24] Y. Zhao, X. Zhang, L. Deng, S. Zhang, *Comp. Chem. Eng.* **2016**, *92*, 37–42.
[25] F. Gharagheizi, P. Ilani-Kashkouli, A. H. Mohammadi, D. Ramjugernath, D. Richon, *Chem. Eng. Sci.* **2012**, *80*, 326–333.
[26] J. O. Valderrama, L. F. Cardona, R. E. Rojas, *Fluid Phase Equilib.* **2019**, *497*, 178–194.
[27] A. Bakhtyari, R. Haghbakhsh, A. R. C. Duarte, S. Raeissi, *Fluid Phase Equilib.* **2020**, *521*, 112662.
[28] http://www.cambridgesoft.com.
[29] https://ilthermo.boulder.nist.gov/.
[30] I. Oprisiu, S. Novotarskyi, I. V. Tetko, *J. Chemom.* **2013**, 5:4.
[31] a) https://cran.r-project.org/; b) L. Breiman, *Machine Learning* **2001**, *45*, 5–32.

Manuscript received: July 1, 2020
Version of record online: November 19, 2020