

Introduction

The numerical representation of a generical chemical structure is obtained with Molecular Maps of Atom-Level Properties (MOLMAPs) technology. [1] This codification technology consists on the pattern of activation of all the structural features of a given chemical system in a trained Kohonen neural network. [2] A structural feature type can be an atom, [3-5] bond [6-8] or an atom-binary inter/intra-molecular interaction. [9, 10] The position of a generical structural feature in a trained Kohonen Neural Network depends on its profile of (combined-)atomic properties. Structural features with similar profiles will activate/will be mapped in nearby neurons. This mapping approach corresponds to an intuitive/generical form of unveiling resemblances and differences between the structural features, of a given type, representing a given chemical system. This codification approach enables the algorithm the possibility to learn non-specific relationships during the training stage that can be transposed to different structural profiles in the validation sets. Additionally, this codification method permits to compare in fixed-size Kohonen matrixes chemical systems of different nature, number of components accounting its molar fractions and respective combinations when dealing with atom-binary inter/intra-component interactions. When a generical chemical system is a molecule, its molar fraction is 1 and a generical structural feature from that molecule will activate a given neuron with the value of 1, this is denominated the Wining Neuron (WN). This neuron corresponds to the lowest Euclidean distance between the structural feature profile of (combined-)atomic properties and the corresponding neuron weights of the trained Kohonen network. The neighbour neurons of the WN may be additionally activated until a predefined level of neighbour neurons, maximum of three levels in this MOLMAP approach, with a progressively lower value going from WN to levels 1N, 2N and 3N. The Kohonen matrix is in fact a torus, considering that the extreme right is linked with extreme left side, and similar approach for the bottom-up sides. It's important to consider that this matrix is transformed into a vector by concatenation of its lines (up-bottom direction).

Experimental setup

To achieve the MOLMAP of a given chemical system different aspects are considered in a sequential form:

- 1- A generical dataset, represented by its smiles strings, is standardized with Chemaxon workframe, [11] specifically the standardizer application with the enabled options 1) mesomerise, 2) add implicit hydrogens and 3) clean 3D, in order for a straightforward comparison of different chemical systems resulting in a standardized .sdf file, (Example: B-TR-std.sdf)

- 2- The resultant .sdf file is submitted to Chemaxon, [11] cxcalc module to calculate the necessary atomic properties and other molecular properties:

```
cxcalc -S -o output_file_name.sdf charge -p 4 -r true -t "pi,sigma,total"
orbitalelectronegativity -p 4 -r true -t sigma atomicpolarizability -p 4
sterichindrance -p 4 -o 2 acc don molecularpolarizability -p 4
molecularsurfacearea -p 4 volume -p 4 -o 2 mass -p 4 input_file_name.sdf
```

The resultant output file example is C-TR-cxcalc.sdf.

- 3- The next stage consists in changing file extension .sdf to .txt for a straightforward reading/gathering of the atomic properties calculated in the previous step. The program built in our labs Atomic-Component-PropertyExtractMarvin (acpextractmarvin.exe), [12] was used in that context with the following command: *acpextractmarvin -n input_file_name.txt*

The obtained output files comprise atomp.txt and component.txt. Our main focus consists on the atomic properties: AtomNumber, qpi (pi charge), qsigma (sigma charge), qtot (total charge), oensigma (orbital sigma electronegativity), pol (atom polarizability), hindrance, acc (number of hydrogen bond acceptor sites), don (number of hydrogen bond acceptor sites). The atomp.txt last column is represented by the label molN_atomN.

Note1: The following stages (4th to 12th) are here explained for assessment of MOLMAPs of atom-binary inter/intra-component interactions of generical molecule as chemical system with the necessary adaptation for determination of MOLMAPs of more complex systems.

Note2: MOLMAPs of atoms of the different chemical system types does not require stages 4, 5 and 7.

- 4- This fourth stage is used in order to obtain atomic-binary inter/intra-component-based MOLMAP. The home-made InterIntra.class application is used to that end. [13] The format and name of the dataset atomp.txt file is changed to E-TR-atomp.csv and used twice (in case of molecule as generical chemical system) in the command line instruction:

```
java      InterIntra      input_component1.csv      input_component2.csv
number_of_structural_features_component1
number_of_structural_features_component2
```

Or in the case of a salt, the cation and anion respective atomp.txt files are converted into two different input files CAT-TR-atomp.csv and AN-TR-atomp.csv and used accordingly in the upper command line instruction.

In case of a more complex chemical system the different combinations of pairs of components are accounted in the command line instruction.

The resultant file in the case of molecule as generical chemical system type corresponds to:

F-Output-InterIntra-TR.txt divided in two different parts in order to be readable and edited in Excel: F-Output-InterIntra-TR-1-1000.txt and F-Output-InterIntra-TR-1001-2082.txt. [14]

5- The following seven combined atomic properties were obtained for each combination of pairs of atoms of a generical molecule:

i) $(-1).qpi(atom1).qpi(atom2)$, ii) $(-1).qsigma(atom1).qsigma(atom2)$, iii) $(-1).qtot(atom1).qtot(atom2)$, iv) $abs[oensigma(atom1)-oensigma(atom2)]$, v) $pol(atom1).pol(atom2)$, vi) $has(atom1).hds(atom2)$ and vii) $hds(atom1).has(atom2)$.

Where has/hbd correspond to hydrogen-bond acceptor and donor sites.

The following files are obtained in this context: [14]

w) G-TR1-1000-PROP_COMB.txt;

x) G-TR1001-2082-PROP_COMB.txt;

6- The next stage comprises the normalization (0-1) of all the (combined)atomic properties, set on training set's maximum and minimum of the corresponding (combined)property, resulting in the example files: [14]

w) H-TR1-1000-PROP_COMB_NORM.txt

x) H-TR1001-2082-PROP_COMB_NORM.txt

7- The obtained files are further processed in order to:

a) Merge combined properties vi) and vii) into a single property: $\max[has(atom1).hds(atom2), hds(atom1).has(atom2)]$.

b) Delete repetitions of identical combination of pair of atoms.

The resultant files comprise: [14]

w) I-TR1-1000-PROP_COMB_NORM_MAX-
ACC1xDON2vsDON1xACC2_WO-REPEAT.txt;

x) I-TR1001-2082-PROP_COMB_NORM_MAX-
ACC1xDON2vsDON1xACC2_WO-REPEAT.txt;

8- Random selection of 29999 combinations of (pairs of)atoms from the TR set (J-KOHONEN.xlsx), [14] convert this selection to a readable format for the Kohonen neural network applet, [2, 15] (J-KOHONEN.txt) [14] and train the network in order to obtain the file, [14] (J-Kohonen-Neural-Network-20x20)

Convert the example files obtained in 7- in a readable format for the Kohonen neural network applet: [14]

- w) J-TR-1-1000-PRE-KOHONEN.txt
- x) J-TR-1001-2082-PRE-KOHONEN.txt

- 9- Submit the previously obtained files w), x) to the trained Kohonen neural network and obtain the resultant example files: [14]

- w) K-TR-1-1000-PRECUTZ-20x20.txt
- x) K-TR-1001-2082-PRECUTZ-20x20.txt

These obtained files result from the Kohonen neural network output and a further change on the first column label to the label molN_atom1_atom2. This edition permits a correct MOLMAP assessment in the next step.

- 10- Submit the previously edited files to cutmapz3N.exe application, [16] (built in our labs), in order to transform the pattern of activation of all the atom(-pairs) within a generical chemical system into its numerical representation in the form of a vector.

Example: 400 positions's vector from a 20x20 Kohonen matrix. The resultant molecule's MOLMAP involves the WN activation value of 1 per atom-pair and its respective value of activation of 0.5 for each of its immediate neighbour neurons N1-N8. The following command is applied in this framework *cutmapz3N input.txt output.txt 26 1 0.5 0 0*. The value 26 stands for the Kohonen matrix dimension 20x20 (20) plus 6 (3+3), the value of added pseudo-dimensions used to account an activation pattern until 3 levels of neighbourhood of a generical WN placed in the 3 nearby levels to extreme sides of the Kohonen matrix (torus). This pseudo-dimension's activation pattern reverts to the respective neurons of the effective Kohonen 20x20 matrix. The values 1 0.5 0 0 represents respectively the WN 1st 2nd 3rd level neighbourhood pattern of activation of a given structural feature.

The resulting example output files are: [14]

- w) L-TR-1-1000-CUTZ-20x20-1-05-0-0.txt
- x) L-TR-1001-2082-CUTZ-20x20-1-05-0-0.txt

- 11- The resulting MOLMAPS are supplementary edited in order to incorporate the respective MOLMAP positions in the first row and the property of interest in the last column (in the case of this example the melting point). This is an amenable form in order to the RF algorithm process a structure property relationship. The edited example file corresponds to: [14]

- y) M-TR-20x20-1-05-0-0.txt

12- The R, [17] version 3.5.3 was used as framework to run the RF algorithm (version 4.6-14).

References

1. Zhang, Q.-Y., Aires-de-Sousa, J. Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* 45, 6, 1775-1783 (2005).
2. <http://neural.dq.fct.unl.pt/jas/jatoon/>.
3. Carrera, G. V. S. M., Nunes da Ponte, M., Rebelo, L. P. N. Chemoinformatic Approaches to Predict the Viscosities of Ionic Liquids and Ionic Liquid-Containing Systems. *ChemPhysChem* 20, 2767-2773 (2019).
4. Carrera, G. V. S. M., Nunes da Ponte, M. Predict the Viscosities of Ionic Liquids and Their Mixtures. *Chemistry-Methods* 1, 214 -22 (2021).
5. Carrera, G. V. S. M., Cruz, M. L., Klimenko, K., Esperança, J. M. S. S., Aires-de-Sousa, J. Prediction of the Phase Composition Profile of Three-Compound Mixtures in Liquid-Liquid Equilibrium: A Chemoinformatics Approach. *ChemPhysChem* 23, e202200300 (2022).
6. Zhang, Q.-Y., Aires-de-Sousa, J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *J. Chem. Inf. Model.* 47, 1, 1-8 (2007).
7. Carrera, G. V. S. M., Gupta, S., Aires-de-Sousa, J. Machine learning of chemical reactivity from databases of organic reactions. *J. Comput. Aided Mol. Des.* 23, 419-429 (2009).
8. Latino, D. A. R. S., Zhang, Q.-Y., Aires-de-Sousa, J. Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* 24, 2236-2244 (2008).
9. Carrera, G. V. S. M., et. al. The Solubility of Gases in Ionic Liquids: A Chemoinformatic Predictive and Interpretable Approach. *ChemPhysChem* 22, 2190-2200 (2021).
10. Carrera, G. V. S. M. The Melting Point Profile of Organic Molecules: A Chemoinformatic Approach. *Adv. Theory Simul.* 5, 2200503 (2022).
11. JChemSuite was used for standardization of chemical structures and calculation of atomic properties, JChemSuite version 19.4.0, ChemAxon, <https://www.chemaxon.com>.
12. 1-MOLMAPs: ACPEXtractMarvin.exe. DOI: 10.6084/m9.figshare.27637707
13. 2-MOLMAPs: InterIntra.class DOI: 10.6084/m9.figshare.27698820
14. Melting Point Profile of Organic Molecules and The Particular Case of 2,4-dichlorophenoxyacetic acid (Reproducibility Datasets)
DOI: 10.6084/m9.figshare.27678402

15. Aires-de-Sousa, J., JATOON: Java tools for neural networks. Chemom. Intell. Lab. Syst. 61, 167-173 (2002).
16. cutmapz3N.exe DOI: 10.6084/m9.figshare.27700074
17. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2022, Vienna, Austria. <https://www.r-project.org/>.