Quantitative Structure Property Relationship Study to Predict Soil Sorption Partition Coefficient

S. Batouche

***Abstract*—**The Koc partition coefficient is an important parameter for determining the distribution of cantaminants in the environment. The experimental value of koc is difficult to measure, that's why it is important to establish models to predict this parameter.

In the literature, several studies have been conducted to predict its value for certain categories of molecules.

In this work, we have undertaken QSPR study to relate Koc to molecular structure. The study was established on a database of 643 molecules. Among the methods used for the establishment of the QSPR model, the multilayer neural networks gave excellent results in comparison with those of the literature.

***Index Terms*—** QSPR- Neural Network RNN- Koc - Linear regression.

I. Introduction

In the field of environment, the soil sorption partition coefficient Koc is an important parameter which provides informations on the behavior, mobility and toxicity of contaminants. Koc is the distribution of chemicals between soils or sediments and water.

The experimental measurement of this coefficient is biased and influenced by several parameters (Temperature T , pH). The experimental measurement of Koc is tedious costly and time consuming. The estimation and prediction of soil sorption partition coefficient by means of models based on the QSPR could be beneficial. QSPR approch has advantages ; there are no expensive experiments and the results are obtained faster.

Several QSPR models predicting Soil sorption partition coefficient have been published [1-3]. The majority of the studies are limited to one specific class of molecules. Moreover, many published models have correlated Koc to molecular descriptors using linear regression based methods such as MLR, ACP and PLS, and SVR method which are not general since the relationship between physico-chemical parameter and molecular descriptors may be nonlinear

The aim of this study is to develop robust QSPR model for predicting the Koc coefficient using a highly heterogeneous data set of 643 chemicals and apply the most recent Neural Network algorithm, the Deep Learning Neural Network (DNN)

II. Material and Methods

A. Data Set

In this study, a data set consisting of 643 chemicals and their experimental koc values were collected from the literature [4-6]. In this work, koc was expressed in logarithmic units and its value in the data set covered nearly 6 log units The distribution of koc in the data set is shown in Figure 1.

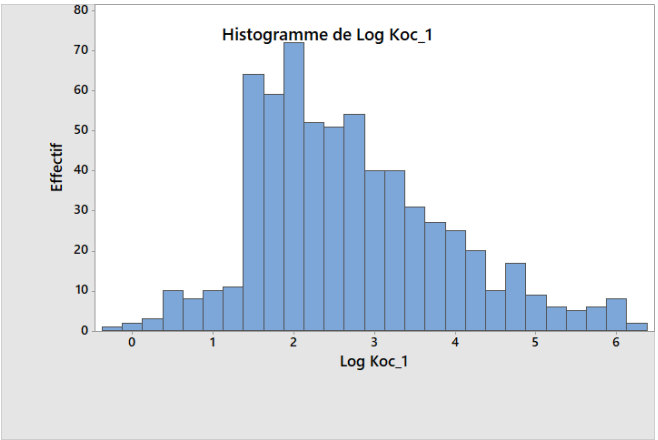

Fig. 1. The distribution of Koc in Data set

The total set was randomly divided into two subsets: a training set for generating the QSPR model and test set for validating the quality of the model.

B. Molecular descriptors

The molecular descriptors were collected from the litterature, P. Gramatica et al [7]. in their work, 1079 descriptors were generated from DRAGON program [8], Many descriptors found to be correlated, they provide the same structural information some of them were excluded to ovoid redundancy in the QSPR model. Genetic algorithm was used to select the most relevant combination of descriptors capable of predicting log Koc.

The selected variables are :

- VP-0: Valence path, order 0
- nHBAcc: Number of hydrogen -bond acceptors
- nAromBond: Number of aromatic bonds

1 University Mohamed Cherif Messaadia, Faculty of Science and Technology, Department des Sciences de la Matière Souk Ahras , Algeria
Laboratoire des Sciences et Tehniques de l'eau et environnement

- MAXDP: Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule).

C. Model development

The objective of this study is to use the Deep Neural Network algorithm for developing a QSPR model for predicting the soil sorption partition coefficient

The deep Neural Network model (DNN) is an Artificial Neural Network (ANN) model composed of many hidden layers which are fully connected. The DNN model can solve very complex problems.

Deep Neural Network (DNN) are composed of many simple computational elements (nodes) interacting across very low bandwidth channels (connections). Nodes in artificial Neural Network are very simple processors inspired by their biological contrepart. The network components are input layer, hidden layer, and out put layer. The Neural Network uses a series of algorithms to detect ralationships in a dataset.

A Deep Neural Network provides a non linear mapping between the inputs (independant variables or descriptors) and the outputs (dependant variables). The architecture of the DNN model is specified by the input layer, output layer, activation function, loss function, optimizer and metric. [9].

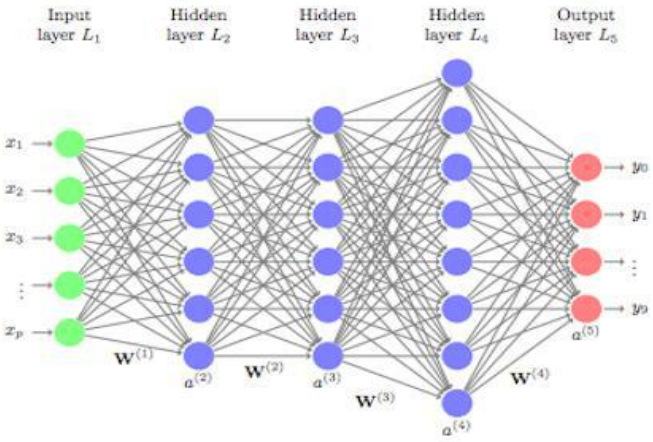

Fig. 2 : Structure of a DNN model [Vishal Yadav]

In this study, the DNN-based QSPR model was developed on the training set using Keras and Tendorflow 2.0 ;

The architecture of the DNN is presented as follows :

- Layer type : Dense
- Number of Hidden Layers : two
- Activation function ; ReLU
- output Layer function: Sigmoid
- Model optimizer : Adam

D. Model validation

In order to verify the ability of the QSPR model to predict external data,, the data set was splited in two subsets:

- A training set (80 % of compounds)
- A test set (20% of compounds)

The test set was used for external validation.

The splitting was performed using a random distribution approach

For evaluating the performance of the DNN based -QSPR model to predict the soil sorption coefficient Koc, two performance metrics : the root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were considered as well as coefficient of determination R^2.

These metrics were calculated as follow :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - Y)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \quad (3)$$

Y_i *represents the experimental* log *Koc value*

$\hat{Y}_i$ *represents the predicted* log *koc values*

n *is the number of compounds*

$\bar{Y}$ *represents the everage of experimental koc values*

III. Results and Discussion

After splitting the Data set into training set (80 % of compounds) and test set (20 % of compounds), 514 compounds were used to train the DNN model while the remaining 129 compounds were used to validate the DNN based QSAR model.

A simple linear regression between experimental and predicted values of Koc was performed and provided the following results.

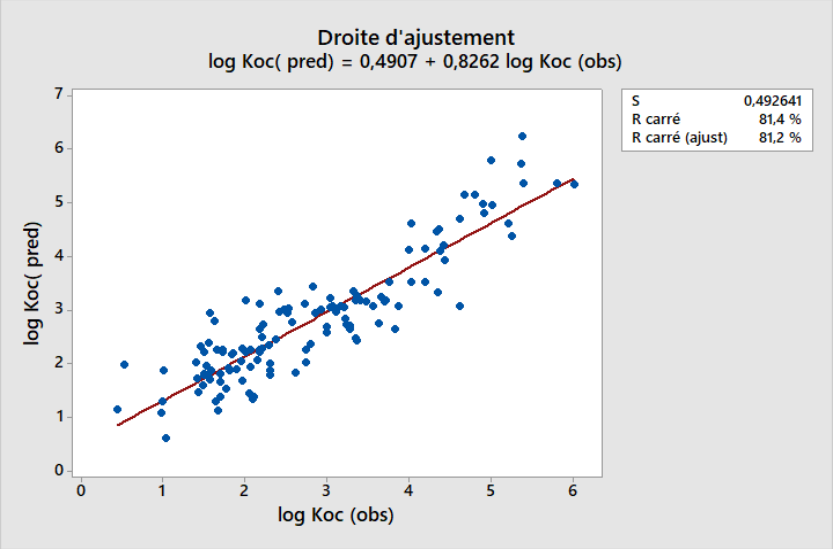

Fig. 3 linear regression for test tcompounds

The performance parameters were calculated, the results are listed in table 2.

TABLE II : The Performance of the model :

Goodness-of-fit			Predictive ability			
n (train)	RMSE (train)	MAE (train)	n(test)	RMSE (test)	MAE (test)	R^2
514	0,2169	0,3610	129	0,2458	0,3834	81,4 %

Figure 4 represents the congergence of RMSE for the training and test set

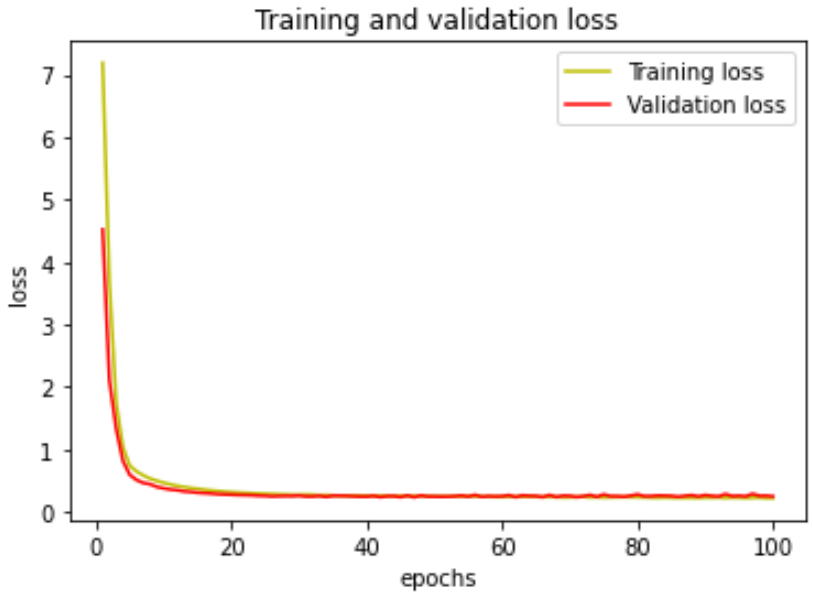

Fig. 4. The convergence of RMSE

The study of linear regression between calculated and experimental log koc values for test compounds leads to value of determination coefficient R^2 higher than 0,6 and low standard error shows the goodness of fit result for the DNN-based QSPR model developed. In addition, RMSE and MAE values which

are less than 0,4 confirm the high predictive abilities of the QSPR model of the QSPR model.

The prediction ability of the QSPR medel obtained in this study was compared with those in previous publishedd studies. The comparison of statistical parameters is shown in table III.

TABLE III : List of published QSAR models of log Koc and statistical parameters :

Study	Algorithm	R^2	RMSE (train)	RMSE (Test)
Shao et al. 2014 [1]	MLR	0,808	0,490	0,475
Gramatica et al. 2007 [7]	MLR	0,820	0,523	0,560
Olguin et al 2017 [3]	MLR	0,809	0,428	0,480
Shao et al. 2014 [1]	SVM	0,817	0,344	0,431
Wen et al. 2012 [10]	MLR	0,790	0,490	0,570
Current study	DNN	0,814	0,2169	0,2458

According to the results in table 3, the DNN-based-QSPR model showed the lowest Root Mean Square Error (RMSE) values compared to the other reported models. Indicateing that the proposed DNN -based-QSPR model developed shows better predictive ability.

IV. Conclusion

In this study, a DNN-based QSPR model was developed on a highly heterogeneous database using molecular descriptors generated from DRAGON program and selected by genetic algorithm-variable subset selection (GA-VSS) software.

The DNN-based QSPR model was trained with 80 % of compounds and validated with 20 % of compounds, the model demonstrates high predictive ability (RMSE = 0,2458 , MAE = 0,3834) and satisfactory goodness-of-fit (R2 = 0,81)

A Comparision between the present model and other models reported in literature, showed that the model proposed is best for prediction of log Koc of chemicals. The QSPR model developed can be used to predict soil sorption partition coefficient (log Koc) of new molecules and will be used for the assessment of the environmental risks of molecules before their development.

The prediction abilities of the QSPR model obtained in this study were compared with those in previous published studies, the results are shown in table 3 :

References

[1] Y. Shao, J. Liu, M. Wang, L. Shi, X. Yao, P. Gramatica , Integrated QSPR models to predict the soil sorption coefficient for a large diverse set of compounds by using different modeling methods, *Atmospheric Environment* vol. 88 pp 212-218, 2014. https://doi.org/10.1016/j.atmosenv.2013.12.018

[2] F. Yang, M. Wang, Z. Wang , Sorption behavior of 17 phthalic acid esters on three soils: Effects of pH and dissolved organic matter, sorption coefficient measurement and QSPR study, *Chemosphere*, vol. .93 pp 82-89, 2013. https://doi.org/10.1016/j.chemosphere.2013.04.081

[3] C. J. M. Olguin, S. C. Sampaio, R. R. dos Reis, Statistical equivalence of prediction models of the soil sorption coefficient obtained using different log P algorithms, *Chemosphere* , vol. 184 pp. 498-504, 2017. https://doi.org/10.1016/j.chemosphere.2017.06.027

[4] A. Sabljic, H. Gusten, H. Verhaar, J. Hermens, QSAR modeling of soil sorption. improvements and systematics of log Koc vs. log Kow correlations, *Chemosphere, vol.* 31, pp. 4489–4514, 1995 https://doi.org/10.1016/0045-6535(95)00327-5

[5] S. Tao, H. Piao, R. Dawson, X. Lu, H. Hu, Estimation of organic carbon normalized sorption coefficient (KOC) for soils using the fragment constant method, *Environ. Sci. Technol*. Vol. 33, pp. 2719–2725, 1999 https://doi.org/10.1021/es980833d.

[6] J. Huuskonen, Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure, *J. Chem. Inf. Comput. Sci.* Vol. 43, pp. 1457–1462, 2003. https://doi.org/10.1021/ci020342j

[7] P. Gramatica , E. Giani, E. Papa, Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, *J.Molecular Graphics and Modelling, vol.* 25, pp. 755–766 ; 2007. https://doi.org/10.1016/j.jmgm.2006.06.005

[8] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON—software for the calculation of molecular descriptors, in: Version 5.3 for Windows, 2005.

[9] N. Wang, D. Zhang, H. Chang, H. Li, Deep learning of subsurface flow via theory-guided neural network, *journal of hydrology* , vol. 584, pp 124700, 2020. https://doi.org/10.1016/j.jhydrol.2020.124700

[10] Y. Wen, L. M. Su, W. C. Qin, L. F. ,Jia He, Y. H. Zhao, Linear and non-linear relationships between soil sorption and hydrophobicity: Model, validation and influencing factors, *Chemosphere*, vol.86, pp. 634-640, 2012. https://doi.org/10.1016/j.chemosphere.2011.11.001