IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

6th National Conference on Science, Technology and Communication Skills – NCSTCS 2K23

Narula Institute of Technology, Agarpara, Kolkata, India

Vol. 10, Special Issue 3, September 2023

To Build a Quantitative Structure Property Relationship (QSPR) Model for a Number of Pesticides to Predict Their Harmful Effect on the Environment Using Machine Learning Methods

Indrani Sarkar[1] and Bibek Dhar[2]

Faculty, Department of Basic Science and Humanities, Narula Institute of Technology, Kolkata, India[1]

Student, Department of Artificial Intelligence and Machine Learning, Narula Institute of Technology, Kolkata, India [2]

Abstract: The properties of a molecule are related to its structure. The objective of quantitative structure-activity/property/toxicity relationship (QSAR/QSPR/QSTR) research is to find out correlation between molecular structures and their biological activities. Models are built by regression analysis using a variety of molecular properties known as descriptors. Some examples of descriptors are: constitutional, topological, geometrical, electrostatic, quantum chemical, molecular orbital (MO) related, thermodynamic and DFT based reactivity descriptors the topological descriptors encode the crucial structural fragments (also known as pharmacophores in case of active drug molecules). The descriptors have direct correlation with activity/property or toxicity data of the molecules. In the present article an attempt has been made to build a QSAR model with 163 pesticides and herbicides to predict the adsorption behavior or mobility of these compounds in soil from their molecular structures. This will give us an insight about the toxicity level of these pesticides and impact on environment and human health.

Keywords: QSAR, Pesticides, Molecular Descriptors, Associative Neural Networks

I. INTRODUCTION

Quantitative structure-property relationship (QSPR) method is used to develop relationships between properties of chemical compounds and their biological/chemical activities. A statistical model is developed which is used to predict the activities of new chemical compounds. The first step in Quantitative Structure Property Relationship (QSPR) is to build the three-dimensional model of the compound.

The structures are taken from different molecular databases like Cambridge database, PubChem, Protein Data Bank, Zinc15 Database and ChEMBL. In the second step several hundred descriptors are calculated for the molecule set by software packages like PaDEL, Codessa Pro, CDK etc. In the third step a few descriptors are chosen from a large pool of descriptors calculated. This is done by a few machine learning methods like ANN and Genetic Algorithm. This selected set of descriptors is small but rich in information which are used to build the QSPR model.

Multiple Linear Regression (MLR) build a correlation model between molecular structure and the experimentally determined activity data through a linear combination of structural descriptors. The numbers of descriptors used to build the model are usually between 3 to 12. Finally, the model is verified by validation some other set of molecules with known biological data are fed into the model and their experimental and calculated (predicted by the model) data are compared. In the present article 163 pesticides have been selected to build a QSPR model. This model will be used to predict the harmful effect of these compounds on environment and animal life.

II. RELATED WORK

The The QSPR models are of two types; Linear and Non-Linear. The linear models (like Multiple Linear Regression) bear a linear relationship between the descriptors and biological property. Nonlinear structure-property relationships produce better results than linear QSPR models because the structure- activity/property relationship is not really linear. Machine learning methods like Artificial Neural Network (ANN) and Genetic Algorithm are employed in building nonlinear models. In computer aided drug design ANN-based tools are applied in pharmacophore mapping, virtual screening, clustering, and modeling of biological activities. A group of indole-based analogues as HIV-1 attachment inhibitors were studied by ANNs [1]. ANNs were also used to study the antimalarial activity of hybrids 4-anilinoquinoline - triazines derivatives with the wild - type and mutant receptor *pf*-DHF (Plasmodium falciparum dihydrofolate reductase) [2].

III. METHODOLOGY

OCHEM is an open-source system that is moderated and managed by users [3]. OCHEM platform consists of two major components: experimental database and the modeling framework. The modeling framework provides variety of machine learning tools for development of QSAR/QSPR models for compounds. Data Preparation: A total of 163 herbicides and pesticides of chloroacetanilide family were selected from the OCHEM Database. These compounds are enlisted in the Hazardous Substances Data Bank (HSDB) and are highly toxic. The experimental values of LogKoc were taken from the database Koc means soil organic carbon absorption coefficient. It is expressed by the equilibrium ratio of the concentration of a compound between soil and water. Koc thus measures the mobility or spreading of a compound in soil. A very high value of Koc of a compound means it is strongly adsorbed into soil and does not spread much throughout the soil. A very low value of Koc means it spreads easily in soil. Koc is used to estimate the spreading of a chemical substance in environment. For pesticides and herbicides compounds with higher Koc is preferred because such compounds have less chance to leach to contaminate ground water. If Koc of a compound is very high (for example, log Koc >4.5), then the substance is likely to have harmful effects on other organisms on ground.

Calculating Molecular Descriptors: The choice of molecular descriptors is one of the most important steps in developing of the successful QSAR model for this model – simple topological EState descriptors and ALogP were selected. Kier and Hall developed the electro topological state (E-state) atom indices in the early 90s [3]. The E-state indices encodes electronic environment of molecular fragments from topology. This is as an important parameter for building QSAR models. E- state indices (2D descriptors) were calculated by OCHEM using E-state program, logP (water solubility) were calculated by ALOGPS program [4,5,6,7,8]

Model Configuration: Out of 5305 descriptors highly correlated descriptors wer excluded and only those descriptors which correlate with the experimental values of Koc were selected by machine learning method like ANN. QSAR modelling require two sets of compounds. One is a Training set for creating (training) a model and a Test set (additional set of compounds that are not used for training) for validation of the new model robustness. Here 129 molecules were taken to train the model and 32 molecules were treated as test set. Validation of the model: The model was validated by using 12 compounds, and the experimental and calculated values of Koc were compared (Table 1).

IV. RESULTS AND DISCUSSION

A scatter plot was plotted to show the measured (experimental value of LogKoc) with the predicted value by the model. (Fig 1). The measured values were plotted along x axis and the values predicted by OCHEM were plotted along y axis. The plot shows the extent of correlation between the measured values and the predicted values of LogKoc. The model basic statistics for each set (correlation coefficient R and cross-validated correlation coefficient q, root mean squared error “RMSE” and mean absolute error “MAE”) are shown in Fig 2. R^2 shows how well data are fit into the model while q^2 is a measure of model stability. The higher those number the better the model. RMSE and MAE represent the difference between the actual values and the predicted values. The smaller those number the better the model. Here the model satisfies the said conditions and shows good statistical parameters.

To validate the robustness of the model 12 compounds with experimental values of LogKoc were fed into the model and the predicted values by the software were compared with the experimental values (Table 1). It was found that the experimental and predicted values match quite well. Fig 3 shows the applicability domain. The Williams plot for standard deviation is shown here. It can be plotted for different p values less than 0.05.

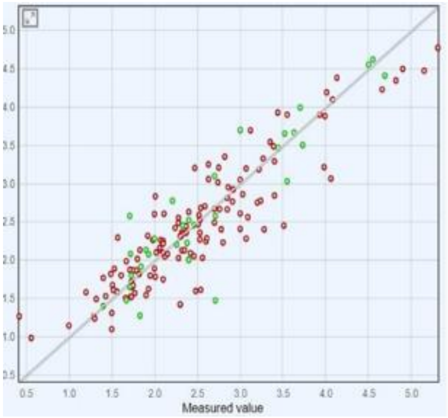
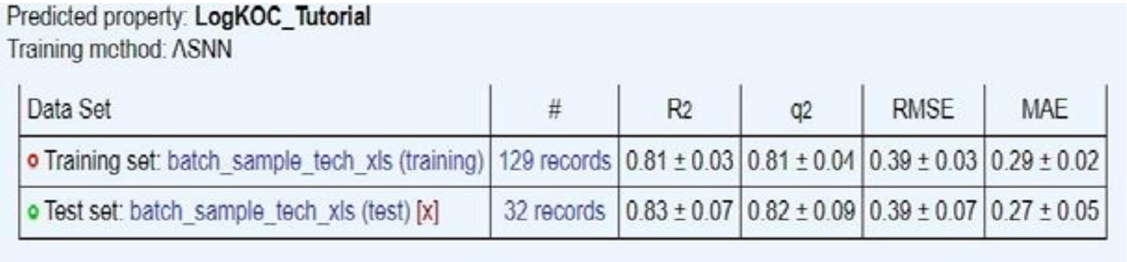


Fig. 1 Scatter Plot



Predicted property: **LogKOC_Tutorial**
Training method: ASNN

Data Set	#	R2	q2	RMSE	MAE
○ Training set: batch_sample_tech_xls (training)	129 records	0.81 ± 0.03	0.81 ± 0.04	0.39 ± 0.03	0.29 ± 0.02
○ Test set: batch_sample_tech_xls (test) [x]	32 records	0.83 ± 0.07	0.82 ± 0.09	0.39 ± 0.07	0.27 ± 0.05

Fig. 2 Statistics of Data Set

TABLE I TWELVE COMPOUNDS FOR VALIDATION

Name of compounds for validation of the model	**LogKoc predicted by ochem**	**ASNN-STDEV**	**Experimental value of LogKoc**
Butralin	3.96	0.42	3.98
Dinitramine	3.67	0.41	3.63
Fluchloralin	3.755	0.34	3.55
Oryzalin	3.081	0.32	3.4
Profluralin	3.878	0.33	4.01
Trifluralin	3.785	0.3	3.93
Aldrin	4.413	0.58	4.69
Chlordane	4.85	0.5	5.15
Endosulfan	4.482	0.5	4.13
Lindane	3.701	0.48	3.0
Methoxychlor	4..771	0.46	4.9
Carbophenothion	4.484	0.32	4.66

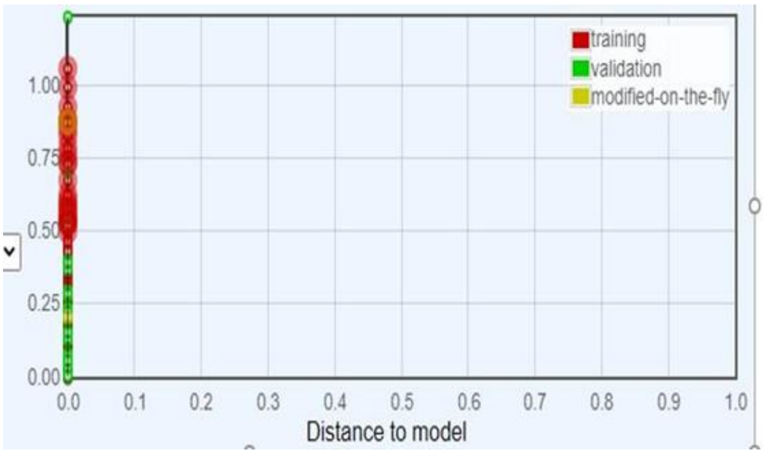


Fig. 3 Applicability domain

V. CONCLUSION AND FUTURE SCOPE

The application of artificial neural networks (ANNs) in QSAR research was first reported in 1971 by Hiller et al. [9]. They tried to classify substituted 1,3- dioxanes as active or inactive with respect to their physiological activity with the use of perceptrons, the early type of artificial neural networks known at that time [10]. Since then, chemists and biologists are using artificial neural networks for predicting different types of biological activities of chemical compounds. There are many software and webservers for QSAR studies. Some examples of 3D QSAR software are the HypoGen module of Catalyst, PHASE, comparative molecular field analysis (CoMFA), and comparative similarity indices analysis (CoMSIA). A few tools for the calculation of molecular descriptors are ADMET Predictor, ChemAxon, PaDEL Descriptor, E DRAGON, CODESSA PRO, PreADMET, ACD/labs and MOPAC. QSAR techniques are two types: linear and nonlinear methods. The linear method uses techniques like linear and multiple linear regressions, partial least squares, principal component analysis and principal component regression. A nonlinear QSAR method includes k-nearest neighbors, artificial neural networks and Bayesian neural nets. A large number of publications are reported on the prediction of physicochemical, ADME, biodegradability, and spectroscopic properties and reactivity of chemical compounds and drugs for recent SARS Coronavirus. Due to their naturalness, clarity and simplicity ANNs are now immensely popular among the scientific communities [9, 10].

REFERENCES

[1]. Hdoufane I. (October 2019). QSAR and molecular docking studies of indole-based analogs as HIV-1 attachment inhibitors. Journal of Molecular Structure, 1193(5), 429-443.

[2]. Hadni H., Elhallaoui M. (August 2019) QSAR studies for modeling the antimalarial activity of hybrids 4-anilinoquinoline-triazines derivatives with the wild-type and mutant receptor pf-DHFR, Heliyon, 5(8), e02357.

[3]. https://www.ochem.eu/home/show.do

[4]. Kier L.B and Hall L.H. (1999). Molecular Structure Description: The Electrotopological State Academic Press: London, 245.

[5]. Hall L.H. and Kier L.B. (1995). Electrotopological state indices for atom types - a novel combination of electronic, topological, and valence state information. J. Chem. Inf. Comput. Sci., 35, 1039– 1045.

[6]. Tetko I.V. and Tanchuk V.Y. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. J. Chem. Inf. Comput. Sci., 42, 1136– 1145.

[7]. Tetko I.V. (2002). Associative neural network.Neur. Proc. Lett. 16, 187– 199.

[8]. Tetko I.V. (2002). Neural network studies. Introduction to associative neural networks. J. Chem. Inf. Comput. Sci. 42, 717– 728.

[9]. Hiller S.A., Glaz A.B., Rastrigin L.A, Rosenblit A.B. (1971). Recognition of physiological activity of chemical compounds on perceptron with random adaptation of structure, Dokl Akad Nauk SSSR, 199, 851–853.

[10]. Minsky M., Papert S. (1969). Perceptrons. MIT Press, Cambridge, MA.