

A simple QSPR model for predicting soil sorption coefficients of polar and non-polar organic compounds from molecular formula

J. Chem. Inf. Comput. Sci.

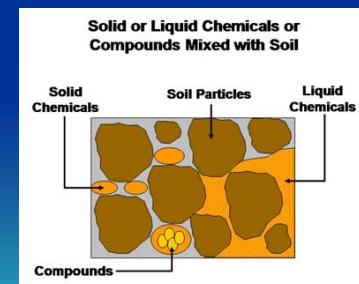
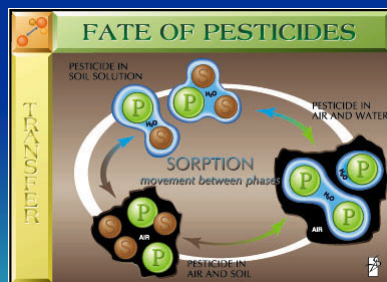
Eduardo J. Delgado, Joel B. Alderete, and Gonzalo A. Jana

(2003)

Presented by:

Mohamed KASSAB

DIAR, Politecnico di Milano



Outlines

- Introduction
- Model description,
- Data set,
- Results,
- Validation set,
- Physical interpretation,
- Conclusion.

Introduction

- The term sorption is used frequently in environmental situations to denote the uptake of a solute by a solid (soil or sediment or component of soil) without reference to a specific mechanism.
- Sorption processes play a major role in determining the environmental fate and impact of organic chemicals.
- Sorption affects a variety of specific fate processes, including:
 - Volatilization,
 - Bioavailability,
 - Biodegradability,
 - Photolysis,
 - Hydrolysis.

- Sorption coefficients, K_{oc} :

- quantitatively describe the organic chemical distribution between an environmental solid, soil, sediment, suspended sediment, and the aqueous phase in contact with at equilibrium

$$K_{oc} = \frac{C_{Soil}}{C_w} \quad (1)$$

C_{soil} is the concentration of solute per gram of carbon in the soil phase, C_w denotes the concentration of solute in the aqueous phase.

The experimental measurement of K_{oc} is expensive, time-consuming and often related with considerable experimental error; consequently, there is a great need for reliable calculation methods which can be used for the prediction of K_{oc} .

Sorption coefficient prediction

There are many approach for predicting sorption coefficient such as:

- Molecular properties that are obtained from quantum chemical calculations.
- Universal solvation model SMx based on semiempirical molecular orbital theory in combination with a dielectrical continuum model.

Since all the above methods, based either on other experimental properties or topological parameters in conjunction with fragment contributions, or quantum chemical calculations, require information and/or computer programs which are not always readily available, Its important to have a model based on **parameters obtained** directly from the molecular formula without any further calculation, especially for scientists with no background in QC.

Model description

- The QSPR model was developed using the Microsoft Windows version of the **Codessa** program which is a chemical multipurpose quantitative activity and structure-property statistical analysis and prediction program, using the best correlation option.
- **Codessa** calculates a total of 38 constitutional descriptors including the following:
 - absolute and relative numbers of atoms, number of single, double, triple, and aromatic bonds,
 - number of rings, number of benzene rings, and molecular weight.

CHEMICAL DATA

- The data set of the sorption coefficients was collected from several literature sources.
- A total of 82 structurally diverse compounds were considered. *
- The data set contains both polar and nonpolar, saturated, unsaturated, aliphatic, aromatic, and polycyclic aromatic compounds covering a log K_{oc} range from about 1 to 6 log units.

* *Syracuse Research Corporation, Physical/Chemical Property Database (Physptop); SRC Environmental Science Center: Syracuse, NY, 1994.*

* *Linders, J. B. H. J.; Jansma, J. W.; Mensik, B. J. W. G.; Otermann, K. National Institute of Public Health and Environmental Protection (RIVM), Bilthoven, The Netherlands, Report No. 679101014.*

82

CHEMICALS



38

**MOLECULAR
DESCRIPTORS**

Best correlation

number of benzene rings,

molecular weight,

number of nitrogen atoms,

number of oxygen atoms,

number of sulfur atoms

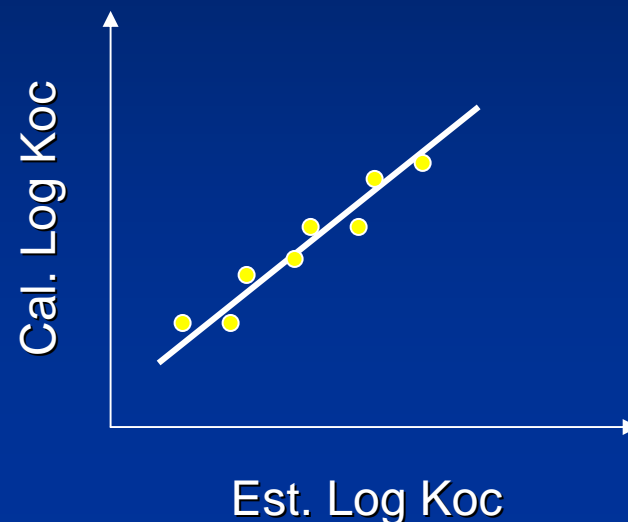
Codessa

QSPR



**PHYSICO-CHEMICAL
PROPERTIES**

Log Koc

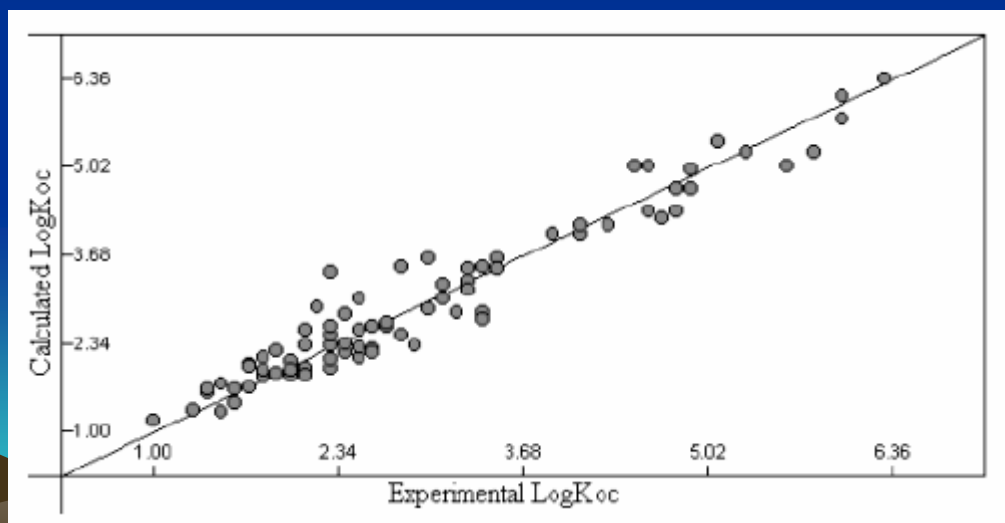


- For the 82 compounds the best correlation equation involving five constitutional descriptors is:

Table 1. Best Five Descriptors Correlation Model of $\log K_{oc}$

descriptor	coefficient	<i>t</i> -test
intercept	0.51 ± 0.16	3.26
number of benzene rings	$0.60 \pm 3.65 \times 10^{-2}$	16.54
molecular weight	$1.01 \times 10^{-2} \pm 5.65 \times 10^{-4}$	17.99
number of N atoms	$-0.48 \pm 6.84 \times 10^{-2}$	-6.95
number of O atoms	$-0.25 \pm 5.12 \times 10^{-2}$	-4.95
number of S atoms	0.61 ± 0.11	5.36

$$\text{Log } K_{oc} = 0.51 + 0.60 N_{\phi} + 1.02 \times 10^{-2} M_W - 0.48 N_N - 0.25 N_O + 0.61 N_S$$



Summary statistics for the correlation of experimental vs calculated log K_{oc} for the 82 compounds are:

$$R^2 = 0.94; \quad (R_{CV})^2 = 0.93; \quad s = 0.33; \quad F = 227.51$$

Where: R is the correlation coefficient, F is the value of the Fisher test, s is the standard deviation of the fit, and RCV is the cross-validated correlation coefficient.

For each data point, the regression is recalculated with the same descriptors but for the data set without this point. The obtained regression is used to predict the value of this point, and the set of estimated values calculated in this way is correlated with the experimental values.

Positive values in the regression coefficients indicate that those descriptors contribute positively to the value of the soil sorption coefficient, whereas negative values indicate that the greater the value of these descriptors the lower the value of K_{oc} .

Model validation

To check the predictive capability of the model, it was tested with an external set of chemicals not included in the training set. The validation data set included 43 compounds with a diverse selection of chemical structures. It contained not only many chemicals that are similar in structure to chemicals in the training set but also some chemicals having other functional groups, namely, alcohols and carboxylic acids.

Statistical performance for calculated vs experimental log *K*_{oc} for the validation was as follows:

$$R = 0.96, R^2 = 0.91, s = 0.09, s^2 = 0.30.$$

These results confirm the predictive capability of the model.

Physical interpretation

In the present model, the **number of benzene rings and molecular weight** are the most important descriptors, for these descriptors indicate the value of K_{oc} goes up as the number of benzene rings and molecular weight increase. This behavior may be explained as follows.

➤ The number of benzene rings encode the hydrophobicity of the compound, i.e., its tendency to exclude from water. Thus, an increase in this descriptor leads to an decrease in the solubility of the compound.

On the other hand, the special mobility of π electrons will result in an enhanced polarizability and interaction energy of unsaturated molecules with the solid and therefore favoring the sorption process. Both phenomena, surely operating simultaneously, lead to an increase in the value of K_{oc} .

➤ The larger the molecular weight the lower the solubility of the compound, leading to an increase in the value of K_{oc} due to its inversely dependence on solubility.

Conclusion

The merit of the QSPR model developed in this article lays in its simplicity. Soil sorption coefficients can be predicted straightforwardly from the molecular formula, for both nonpolar and polar compounds, without any calculation of indices or more complicated quantum chemical calculations or without the need of having the values of polar fragment contributions. Therefore the model is applicable to compounds with polar fragments for which no group contributions have been fitted before. Currently, the model is being developed to include organophosphorus aliphatic and aromatic compounds.