The InChI-based optimal descriptors:
QSAR modeling of toxicity of organometallic compounds

A.A. Toropov, A.P. Toropova and E. Benfenati

Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy

Abstract

Quantitative structure - activity relationships (QSAR) for toxicity toward rats (-pLD50) have been built by means of the optimal descriptors. Comparison of the optimal descriptors calculated with the International Chemical Identifier (InChI) and the optimal descriptors calculated with simplified molecular input line entry system (SMILES) has shown that the InChI-based models give more accurate prediction for the above-mentioned toxicity of organometallic compounds. These models were obtained by means of the balance of correlation: one subset of the training set (subtraining set) plays role of the training; the second subset (calibration set) plays role of the preliminary check of the models. We found that the balance of correlations is more robust for the toxicity thanthe classic scheme (training set – test set: without of the calibration). Three splits into the subtraining set, calibration set, and test set were examined.

Method

Dataset

The numerical data on the toxicity towards rats (LD50, in mg/kg, oral exposure) for 56 compounds were taken from U.S. Library of Medicine [1]. The modeled endpoint is -p(LD50). These compounds contain Ni, Na, Cu, As, Fe, Ag, Mg, Mn, Co, K, Au, Sb. The metal atoms are present as single or multiple atoms of the same or different elements in the same compound. The metal is present in different oxidation state for the same element in the heterocyclic and aromatic ligands which contain oxygen, nitrogen, and sulphur. Thus the set is very diverse.

SMILES notations have been generated with the ChemSketch software [2]. The work set (n=56) was randomly split into a subtraining (n=23), calibration (n=23) and a test set (n=10). Three different splits have been examined. These splits are random, but ranges of the log(LD50) are similar for all subtraining sets, all calibration set, and all test sets.

SMILES-based optimal descriptors

The descriptors used in this study were calculated with SMILES attributes. The SMILES attribute is a combination of SMILES elements. The SMILES element is a group, that contains four, two, or one symbol of a SMILES notation.

In this study 17 SMILES elements have been used: three elements of four symbols are “+2”, “+3”, “@@”, “Ag”, ”As”, “Au”, “Cl”, “Co”, “Cu”, “Fe”, “O=”, “Mg”, “Na”, “Mn”, “Ni”, “Sb”; and 46 elements of one symbol are “#”, “ (“, “ +”, ” –“, “.”, ”/” ,”1” ,”2”, ”3”, ”4”, ”5”, ”=”, ”@”, ”C”, ”F”, ”H”, ”K”, ”N”, ”O”, ”P”, ”S”, ”[“, ”\”, ”c”, ”n”, ”o”, and ”s”.

Optimization of the correlation weights by the Monte Carlo method. Two systems of the modelling were examined: 1) the maximization of the correlation coefficient between the DCW(LimN) and log(LD50) for the subtraining set and calibration set (The classic scheme); 2) the maximization of the criterion calculated as the following:

BC = Rs + Rc – ABS(Rs-Rc) * 0.1 (1)

where Rs and Rc are correlation coefficients between the DCW(LimN) and log(LD50) for the subtraining set and calibration set, respectively (balance of correlations).

The SMILES-based optimal descriptor is calculated as the following:

DCW(LimN) = Σ CW(1SAk) + Σ CW(2SAk) + Σ CW(3SAk) (2)

where CW(1SAk), CW(2SAk), and CW(3SAk) are the correlation weights for the above-mentioned SMILES attributes. The LimN is a parameter of the model used to classify the SMILES attributes into two categories: rare or not rare. Our hypothesis is “rare SMILES attributes are able to lead to overtraining”. The influence of the rare attributes may be blocked, if correlation weights of the rare attributes are fixed equal to zero.

InChI-based optimal descriptors

InChI-based optimal descriptors have been studied with the same three splits into subtraining set, calibration set, and test set. The InChI for organometallic compounds have been constructed with the ChemSketch software. The InChI-based optimal descriptors are calculated as the following

DCW(LimN) = ∑ CW(IAk) (3)

where IAk is an attribute of the InChI.

The InChI attributes used for calculation of the DCW with Eq. 2 are the following: “(10”, “(11”, “(12”, “(13”, “(14”, “(15”, “(16”, “(17”, “(18”, “(19”, “(20”, “(21”, “(22”, “(23”, “(24”, “(25”, “(27”, “ (28”, “(2”, “(3”, “(4”, “(5”, “(6”, “(7”, “(8”, “(9”, “(“,” *”,” +2”,” +3”, “+”, “,10”, “,11”, “,12”, “,13”, “,14”, “,15”, “,16”, “,17”, “,18”, “,19”, “,20”, “,22”, “,23”, “,24”, “,25”, “,26”, “,28”, “,29”,”,1”, “,2”, ”,3”, “,4”, “,5”, “,6”, “,7”, “,8”, “,9”, “,”, “-10”, “-11”, “-12”, “-13”, “-14”, “-15”, “-16”, “-17”, “-18”, “-19”, “-20”, “-21”, “-22”, “-25”, “-27”, “-28”, “-29”, “-30”, “-1”,”-2”, “-3”, “-4”, “-5”, “-6”, “-7”, “-8”, “-9”, “-“, “.”, “/p-2”, “/”, “0”, “1”, “2”, “3”, “4”, “5”, “6”, “7”, “8”, “9”, “:;:;” “:;:” “:;” “:” “.” “=”, “Ag”, ”As”, “Au”, “C10”, “C11”, “C13”, “C14”, “C15”, “C16”, “C17”, “C18”, “C24”, “C2”, “C3”, “C4”, “C5”, “C6”, “C7”, “C8”, “C9”, “Cl”, “Co”, “Cu”, “C”, “Fe”, “F”, “H10”, “H11”, “H12”, “H14”, “H15”, “H16”, “H17”, “H19”, “H20”, “H34”, “H36”, “H2”, “H3”, “H4”, “H5”, “H7”, “H8”, “H”, ”I”, “K”, “Mg”, “Mn”, “N2”, “N3”, “N4”, “N5”, “Na”, “Ni”, “N”, “O2”, “O3”, “O4”, “O5”, “O6”, “O7”, ”O8”, “O9”, “O”, “P”, “Sb”, “S”, “b”, “c”, “h10”, “h13”, “h1”, “h2”, “h3”, “h5”, “h7”, “h8”, “h”, “I”, “m11”, “m1”, “m”, “n”, “p”, “q”, “s1”, “t10”, “t13”, “t2”, and “t7”

SMILES notations are a convenient representation from the chemical point of view. InChI representation is not transparent. However, InChI is more informative format for the QSAR modeling of the toxicity of organometallic compounds by the optimal descriptors (Figs 1 and 2). The balance of the correlations was used for the QSAR modeling of the toxicity of organic compounds.

The second probe of the Monte Carlo optimization for the DCW(16) for InChI-based optimal descriptors gave the following model:

-pLD50 = 2.0415 - 0.2187 *DCW(16) (4)

n=23, r2=0.710, s=0.557, F=52 (subtraining set);
n=23, r2=0.680, s=0.541, F=45 (calibration set);
n=10, r2=0.803, s=0.481, F=33 (test set)

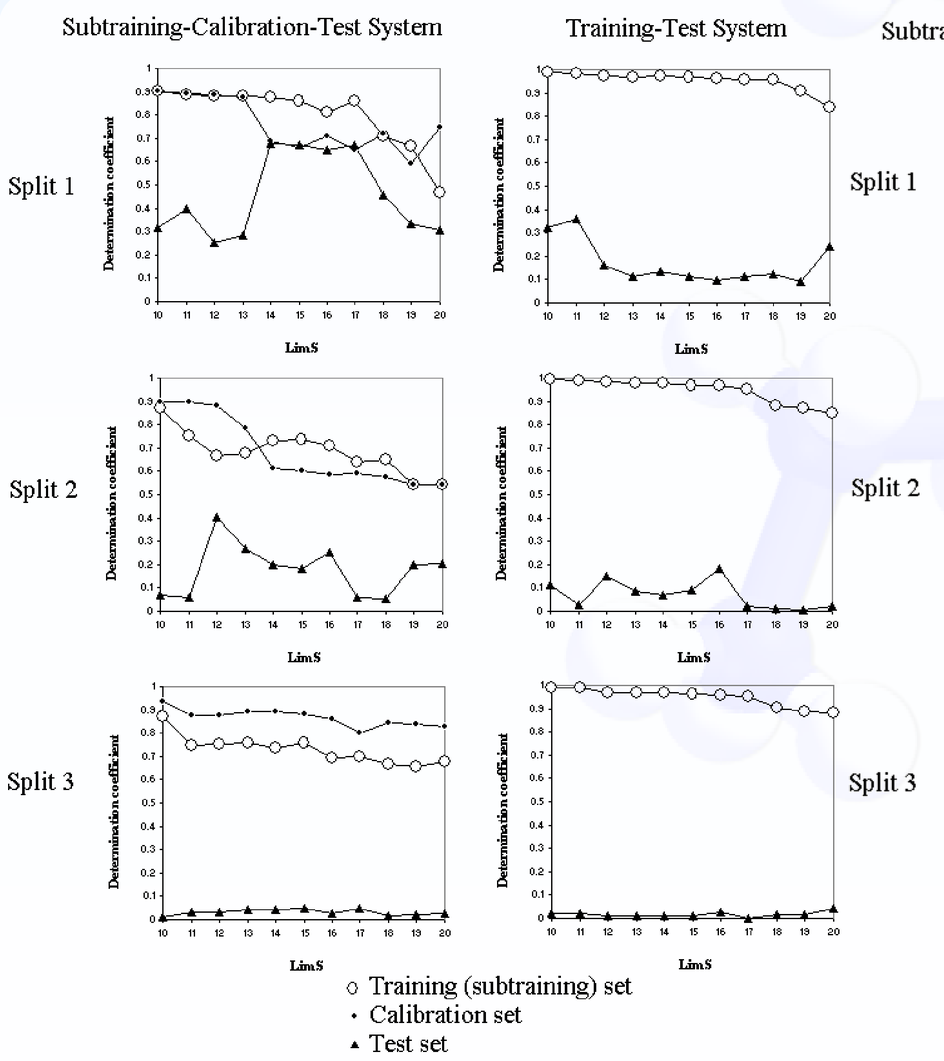


Figure 1. **SMILES-based models**

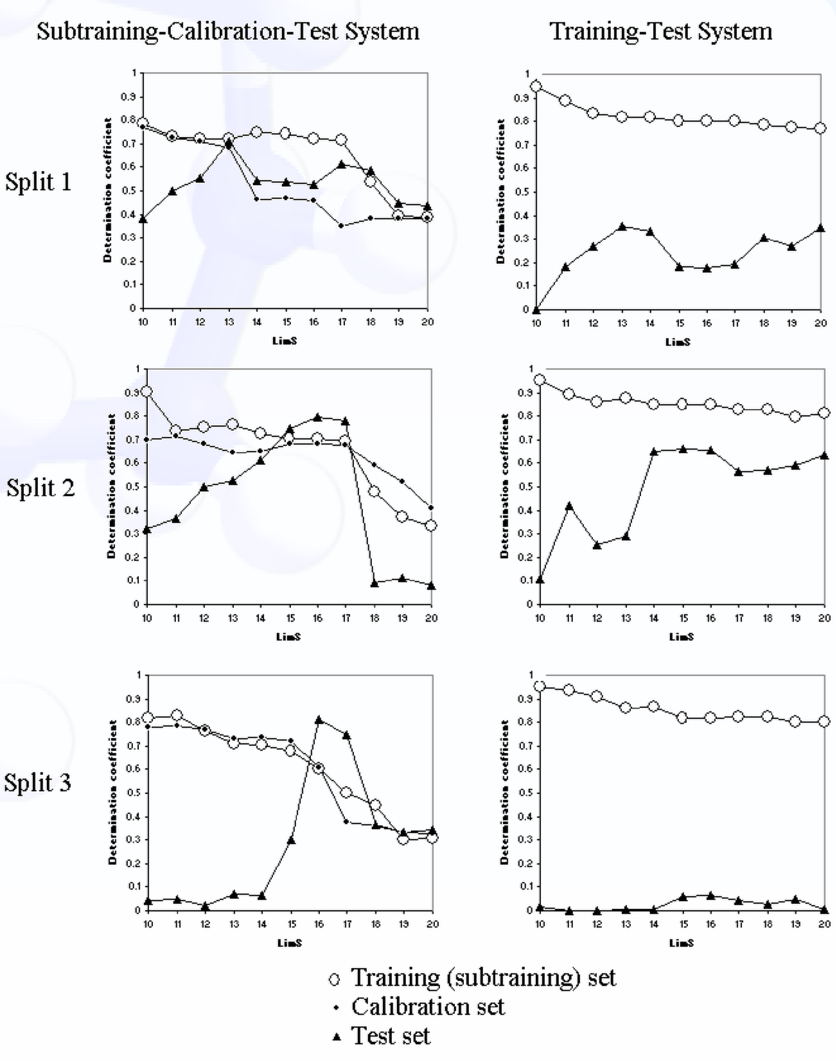


Figure 2. **InChI-based models**

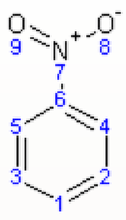
Results and Discussion

One can see from Figs 1 and 2 that optimal InChI-based optimal descriptors give better models than the SMILES-based descriptors. The Subtraining-Calibration-Test System (the correlation balance) is better predictor for the toxicity of organometallic compounds than more traditional Training-Test System.

SMILES format has transparent interpretation for each symbols (c, C, =, #, etc).

InCh format (Fig.3) is a more complex system. However, being aimed to provide unique representation of the molecular system, the format may be useful and interesting alternative of SMILES and molecular graph in the QSPR/QSAR analyses [3].

InChI=1/C6H5NO2/c8-7(9)6-4-2-1-3-5-6/h1-5H



Description	Layers
Formula	C6H11NO2
Connectivity	8-7(9)6-4-2-1-3-5-6
Hydrogen atoms	h1-5H

Figure 3.

Example of the InChI layers

References

[1] U.S. Library of Medicine (2008) <http://toxnet.nlm.nih.gov/>
[2] ACD/ChemSketch Freeware, version 11.00, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2007
[3] Toropov, A.A., Toropova, A.P., Benfenati, E., Leszczynska, D., Leszczynski, J. J.Math. Chem. In press.

Conclusions

1. InChI-based optimal descriptors gave more accurate prediction for the toxicity toward rats of the organometallic compounds than the SMILES-based optimal descriptors; the SMILES-based descriptors gave satisfactory prediction for the split A (only), whereas the InChI-based descriptors gave satisfactory prediction for all three splits; **2.** The balance of correlations (i.e., training with two sets: the subtraining set and calibration set) gave more accurate prediction than the classic scheme (i.e., modeling by scheme of “training-test”, without the calibration); **3.** The split into the subtraining set, calibration set, and test set have considerable influence on the predictive ability of the optimal descriptors.

Acknowledgements

FEDERCHIMICA
AISPEC
Associazione nazionale imprese chimica fine e settori specialistici

The authors thank the *Marie Curie Fellowships* for financial support, through the contract MIF1-CT-2006-039036 - CHEMPREDICT, the EC project CAESAR (contract SSPI-022674), FEDERCHIMICA and Filomena Tarsitano (Lamberti S.p.A.)

CAESAR
ChemPredict