# Ensemble approach to applicability domain of SAR/QSAR models.

**N. Kireeva, G. Marcou, A. Varnek***
*Laboratoire d'Infochimie, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg, 67000, France*

**I. Baskin**
*Moscow State University, Moscow, Russia*

**K. Migita, M. Arakawa, K. Funatsu***
*University of Tokyo, Japan*

Efficient utilization of SAR/QSAR models in virtual screening is closely related to *applicability domain* (*AD*) approach which controls whether a given model can be applied to a query molecule. Here, we systematically investigated an impact of different *AD* approaches on accuracy of prediction of stability constants (log$K$) of the 1:1 complexes of metals ($Ba^{2+}$, $Eu^{3+}$) with organic molecules in water, on one hand, and on the coverage of the chemical space (the percentage of rejected compounds), on the other hand. The regression models were built on the ISIDA's fragment descriptors  using Support Vector Machine method and validated in external 5-fold cross-validation procedure. The following AD approaches have been used:

- distance based (Z-kNN) proposed by Tropsha *et al.*  and adopted to ISIDA ,
- range based (or bounding box), in which the program rejects a query compound if one of descriptors is outside of its *Min – Max* range calculated for the training set;
- fragments based (*Fragment Control* and *Common Fragments*) in which the program controls the type of particular fragments and/or their relative populations in query compound compared to those parameters obtained for the training set.
- one class SVM classification model (SVDD),
- two class SVM classification model. In this case, two different data sets - extracted from the NCI database  and complexants of $Cu^{2+}$ - have been used as negative examples.

Our results show that neither of examined *AD* approaches is sufficient to reject all compounds which are chemically too different from those in the training set. However, a combination of different *AD*s allows user to filter properly the screened database and to provide reasonable compromise between the accuracy of prediction and the coverage of chemical space.

**References.**