AlogPS (Aqueous solubility and Octanol/Water partition coefficient)

Dataset profile

The model predict octanol/water partition coefficient (logP) and solubility in water (logS). Both these parameters are important important for drug discovery. The model is further development of ALOGPS 2.1 program [Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, *J. Chem. Inf. Comput. Sci.*, **2002**, 42, 1136-45 and *Tetko, et al Estimation of aqueous solubility of chemical compounds using E-state indices, J. Chem. Inf. Comput. Sci., **2001**, 41, 1488-93*] which is available at Virtual Computational Laboratory (VCCLAB) site. This program was assessed in several benchmarking studies and was top-ranked for prediction of *in house* Pfizer and Nycomed [Mannhold et al, Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. J Pharm Sci. 2009 Mar;98(3):861-93. doi: 10.1002/jps.21494.].

The data for logP and logS were taken from these two previous publications as well as were merged with those collected at OCHEM web site. The training sets included 16647 and 6778 unique compounds for logP and logS properties, respectively. The data were filtered from the outliers using an automatic p-value based filtering feature of OCHEM (article in preparation). Considering high inter-dependency of both properties, there were modeled simultaneously, using multi-learning feature of OCHEM [Varnek et al, Inductive transfer of knowledge: application of multi-task learning and feature net approaches to model tissue-air partition coefficients. J Chem Inf Model. 2009 Jan;49(1):133-44. doi: 10.1021/ci8002914] to increase the applicability domain of the models.

Data preprocessing

All chemical structures were processed using OCHEM cleaning and standardization protocols.

Descriptors

This model was built using EState descriptors (electrotopological EState indices) using program developed by Dr. Tanchuk, which was also used to develop ALOGPS 2.1 model.

Validation

The model was built using 5-fold cross validation.

Statistical parameters

Prediction accuracy

The basic prediction accuracy parameters according to the 5-fold cross-validation procedure are:

Property	# records	RMSE	MAE	R^2	r_2 (Coefficient of determination)
logP	16912	0.42	0.30	0.95	0.95
logS	8102	0.70	0.52	0.90	0.90

Applicability domain

The prediction accuracy is estimated using ASNN-STD. This distance to model was shown to provide the best assessment of the accuracy of prediction as described in [Tetko et al, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, J Chem Inf Model. 2008 Sep;48(9):1733-46. doi: 10.1021/ci800151m].