

Cytochrome P450 – Drug interaction

Two QSAR datasets were used to develop classification models able to predict the drug interaction with two of the most important isoform of Cytochrome P450 involved in the drug metabolism, CYP3A4 and CYP2C9. The activity data for both the isoforms were derived from the CYP bioactivity database developed by the National Institutes of Health Chemical Genomics Center (NCGC), which collects 17,143 drug-like compounds with the activity outcome for five CYP isoforms. The data curation procedure allowed to obtain a shared set composed by 9,122 molecules with activity data for both the considered isoforms and two external sets which includes 2,996 and 2,818 molecules for CYP3A4 and CYP2C9, respectively. Each external set comprises the molecules with annotated activity class for one isoform only.

The shared set was randomly split into a training set of 6,385 compounds (70%) and a test set of 2,737 compounds (30%), keeping the active/inactive proportion of both the isoforms (49:100 and 66:100 for 2C9 and 3A4, respectively).

For each isoform the excel file is composed by three sections: training set, test set and external set. The compounds are identified by the SMILES and described by the activity class and the selected Dragon molecular descriptors.

For further details on performance of the QSAR models, molecular descriptors and their interpretation, please have a look to the referenced scientific paper.

CYP3A4_Dataset

Excel file contains data for 12,118 drug-like compounds organized in three sections:

- Training set (6,385 compounds);
- Test set (2,737 compounds);
- External set (2,996 compounds).

For each compound are provided the class of activity and 8 molecular descriptors: nBM, nBO, C%, SIC5, ATSC4p, ATSC6i, nROH, NaasC.

CYP2C9_Dataset

Excel file contains data for 11,940 drug-like compounds organized in three sections:

- Training set (6,385 compounds);
- Test set (2,737 compounds);
- External set (2,818 compounds).

For each compound are provided the class of activity and 10 molecular descriptors: Sp, nBM, ARR, nPyrimidines, HyWi_B(m), GATS2i, Eta_betaP_A, nRNR2, F01[C-N], MLOGP.

Conditions

The datasets are freeware and may be used if proper reference is given to the authors. Please, refer to the following paper:

Nembri, S.; Grisoni, F.; Consonni, V.; Todeschini, R. In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. *Int. J. Mol. Sci.* 2016, 17, 914.