Instructions for the demo standalone version of ALOGPS 2.1

The program calculates logP=log$_{10}$(octanol/water) and logS=log$_{10}$(mol/L). It uses SMILES, sdf (MDL) and mol2 (Sybyl) files (two last formats are available in the batch mode and are recognized using file extensions ".sdf" and ".mol") as input data. This program (Tetko et al. 2001, Tetko and Tanchuk 2002) was developed using Associative Neural Network (ASNN) method (Tetko 2008). Benchmarking results of this algorithm were published at (Mannhold et al. 2009, Tetko et al. 2009a, Tetko et al. 2009b, Tetko et al. 2016).

The program has **LIBRARY**, **SINGLE MOLECULE and BATCH** modes. You can proceed to the next mode by pressing <enter>.

General information

Only neutral molecules are used in ALOGPS 2.1. You can also provide structures like C[NH2+]C.[Cl-] that will be automatically converted to C[NH2](Cl)C. In the case if molecule has total positive (or negative) charge, a number of [Cl-] (or [NH4+]) will be automatically added until the molecule will become neutral one. Thus if you provide the molecule as C[NH2+]C, the same result will be calculated.

Chlorides, bromides and iodides of molecules should be made available as one continuous fragment with atom of nitrogen with valence +5. For example, isoprenaline hydrochloride should be represented not like a mixture of isoprenaline and chloride acid, OC(C1=CC(O)=C(O)C=C1)CNC(C)C.[H]Cl but as c1cc(O)c(O)cc1C(O)C[NH2](Cl)C(C)C, i.e. using the same representation that is adopted in KOWWIN. Of course, if this molecule will be provided as Oc1ccc(C(O)C[NH2+]C(C)C)cc1O.[Cl-] or as Oc1ccc(C(O)C[NH2+]C(C)C)cc1O it will be correctly converted to c1cc(O)c(O)cc1C(O)C[NH2](Cl)C(C)C. However, OC(C1=CC(O)=C(O)C=C1)CNC(C)C.[H]Cl will produce an error. Of course, isoprenaline (logP=0.05 logS=-1.57) and isoprenaline chloride (logP=0.23, logS=-2.58) have quite different physicochemical parameters and that is why so important to provide proper structures for the analysis.

Notice, that, for example, CN(Cl)C (logP=-1.22, logS=0.72) and C[NH2](Cl)C (logP=-2.04, logS=-1.06) are completely different molecules from the point of view of ALOGPS.

The analyzed molecules should have at least one atom of carbon (in general program predicts reliably molecules with at least 3-4 carbons). The program will not crash (should not!) if there is an error in SMILES. It will report an explanatory message about the error. The program will not try to correct even apparent SMILES errors.

SINGLE MOLECULE MODE:

The diversity of training sets used in logP and logS programs was very different. Whenever a new molecule contains index that was not used in the training set (or all molecules with such index were outliers), a warning message will be printed. For example, if you try to analyse carbon dioxide O=C=O, both logP and logs program will indicate that:

LOGP:
UNRELIABLE logP: some molecular indices were missed in our training set.
LOGS:
UNRELIABLE logS: some molecular indices were missed in our training set.

In the case if molecule was positively charged, the next message will be also printed:

UNRELIABLE values: charged molecule was made neutral by adding .[Cl-] group.

ALOGPS program has been developed only with C, N, O, S, P, F, Cl, Br and I. All other atoms will provide UNRELIABLE estimations by this program.

If some compounds are provided in the **LIBRARY MODE** and if they are used to estimate properties of the analysed molecule, such compounds will be shown together with their experimental values and r^2 (Tetko 2002, Tetko and Tanchuk 2002). The nearest neighbours are always reported for molecules from the aqueous solubility set.

For example, the ALOGPS will produce the next output for analysis of CCCCC:

```
logP=3.41   logS=-2.53
```

```
logP knn=96 sigma=0.99
logS knn=26 sigma=0.64 similar molecules:
 -3.18  the_same    CCCCC
 -3.84  r*r=0.62    CCCCCC
 -2.64  r*r=0.51    C1CCCC1
 -3.10  r*r=0.46    C1CCCCC1
 -4.53  r*r=0.44    CCCCCCC
 -2.73  r*r=0.42    CCCCCCl
 -2.03  r*r=0.41    ClCCCC
 -2.63  r*r=0.37    CC(Cl)CCC
```

The first line indicates logP and logS values predicted using ASNN. The second and third lines indicate parameters of ASNN. All following lines indicate experimental values, r*r, square of Spearman rank correlation coefficient that measures similarity of the detected and analysed molecule in the space of models (see (Tetko 2002, Tetko 2008) with description of the ASNN method), SMILES of the detected molecules.

BATCH MODE

The input data format is SMILES codes. Use one SMILES per line, e.g.:

```
CC     1.81
CI     1.51
BrC    1.19
C=C    1.13
C#C some comments about this molecule

ClC    0.91
FC     0.51
```

Only the first lexeme will be analysed and the rest of the line will be skipped. After the SMILES code you can put any comments, experimental values, etc. You can also insert empty lines in the file.

The calculated values, if they are not reliable, will be marked. For example, for the carbon dioxide the program will indicate

```
-0.63  0.61   logP   logS
```

i.e., both logP and logS results are not reliable. This result is calculated because this molecule is very specific and contains E-state index Se2C2O1d (atom C connected by double bonds with two oxygen) that is observed in our data set only once.

The calculated results of the batch mode are stored in the **result.txt** file. The files **good_sm.txt** and **bad_sm.txt** store all SMILES that were processed and not processed, respectively.

LIBRARY MODE

This mode provides the user's training feature of ALOGPS 2.1 program. The values of the user's compounds will be (if required) automatically used to improve predictions of new molecules. The input data format should be SMILES followed by logP or logS (in $\log_{10}$(mol/L)) data. The use of the user's compounds in ALOGPS program can dramatically improve prediction ability of the program (Tetko and Poda 2004, Tetko et al. 2008, Tetko et al. 2016).
Warning: before adding a library it is suggested to analyse all compounds in a BATCH mode and to use only compounds that are error free. See also **library.pdf** file with a description of the LIBRARY mode.

The program processes more than 10,000 compounds per minute (depending on the computer). Let us know your comments.

The on-line demo version of ALOGPS 2.1 is available at **http://www.vcclab.org/lab/alogps.**

References

Mannhold, R., et al. (2009). "Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds." J. Pharm. Sci. **98**(3): 861-893.

Tetko, I. V., et al. (2001). "Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices." J. Chem. Inf. Comput. Sci. **41**(5): 1407-1421.

Tetko, I. V. (2002). "Neural network studies. 4. Introduction to associative neural networks." J. Chem. Inf. Comput. Sci. **42**(3): 717-728.

Tetko, I. V. and V. Y. Tanchuk (2002). "Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program." J. Chem. Inf. Comput. Sci. **42**(5): 1136-1145.

Tetko, I. V. and G. I. Poda (2004). "Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds." J. Med. Chem. **47**(23): 5601-5604.

Tetko, I. V. (2008). "Associative neural network." Methods Mol. Biol. **458**: 185-202.

Tetko, I. V., et al. (2008). "Calculation of lipophilicity for Pt(II) complexes: Experimental comparison of several methods." J. Inorg. Biochem. **102**(7): 1424-1437.

Tetko, I. V., et al. (2009a). "Accurate In Silico logP Predictions: One Can't Embrace the Unembraceable." QSAR Comb. Sci. **28**(8): 845-849.

Tetko, I. V., et al. (2009b). "Large-scale evaluation of log P predictors: local corrections may compensate insufficient accuracy and need of experimentally testing every other compound." Chem. Biodivers. **6**(11): 1837-1844.

Tetko, I. V., et al. (2016). "Prediction of logP for Pt(II) and Pt(IV) complexes: Comparison of statistical and quantum-chemistry based approaches." J. Inorg. Biochem. **156**: 1-13.