Training the logP Plugin

This manual gives you a walk-through on how to train the log*P* Plugin:

Introduction

If you think your experimental data could improve the performance of the default log*P* calculator, you can take advantage of the supervised log*P* learning method that is built into the logP calculator.

If you create a local log*P* model, the scope of the log*P* calculator will be limited. It means that the calculated log*P* will only provide reasonable prediction for a few types of structures. Practically only those types of structures will be predicted correctly which were introduced to the training set during the teaching process. For example, if the training set contains only certain types of hydrocarbon but no other functional groups are present in the training set, the predicted log*P* of any amine-like molecule will not be accurate.

Therefore you need to be aware that a more robust general log*P* model requires a large, diverse training set with thousands of structures. You can generate a log*P* training library with the cxtrain command line tool.

Training steps

Preparing the input file

As the first step of the training you have to create a training set from your experimental data. The training set should have a format which supports saving molecular properties (SDF or MRV). This can be easily done by using the graphical user interface of Instant JChem. This training set must contain the following items:

- structure
- logP values in a property field named LOGP

See the following bit of an example file as an example (logP_trainingset.sdf):

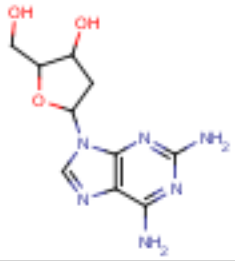
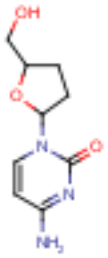
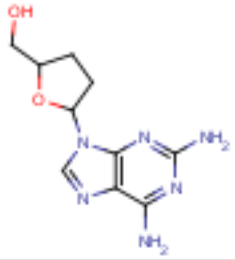
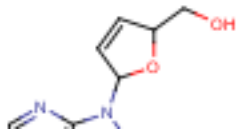
CdId	Structure	logP
403		-0.52
404		-1.30
405		-0.46
406		-0.35

Fig. 1 Example file used for training

Creating the training library

Then you have to run the training algorithm which creates a log*P* training library from your pre-compiled set. Execute the following command from command line:

```
cxtrain logp -t LOGP -i [library name] -a [training file]
```

Example

```
cxtrain logp -t LOGP -i mylogp -a logP_trainingset.sdf
```

The created log*P* training library *mylogp* can be used via marvinsketch, cxcalc, or chemicalterms.

Applying the training library

MarvinSketch

To apply the pre-generated training library *mylogp* in MarvinSketch, do the following steps:

- Choose MarvinSketch menu *Tools > Partitioning > logP*.
- Select the *User defined* method to activate the training option.
- If you have created multiple training sets, choose one from the dropdown list below the checkbox.

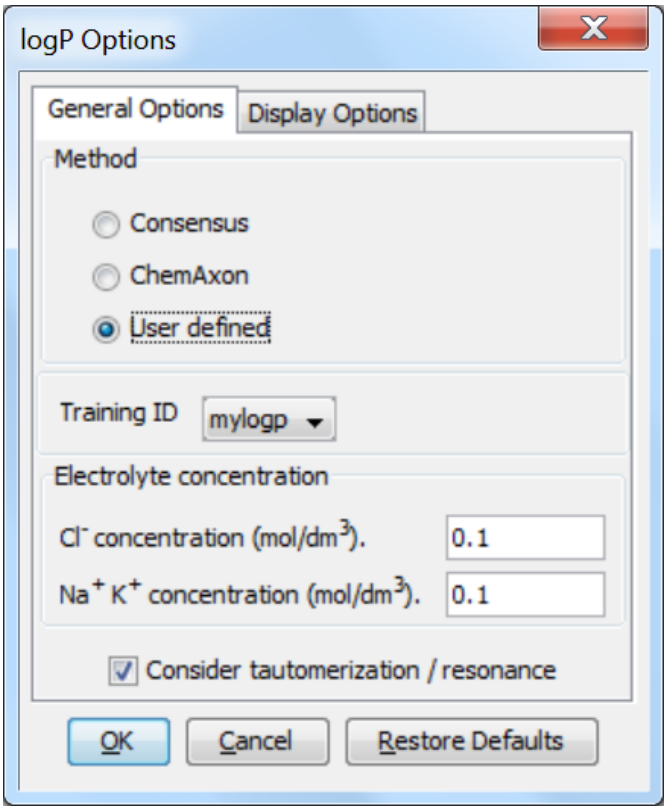


Fig. 2 The log*P* options window showing how to apply the training library

Cxcalc

To apply your log*P* dataset use the *--trainingid* and the *--method* parameter:

```
cxcalc logp --method user --trainingid [library name] [input file
/string]
```

Example

```
cxcalc logp --method user --trainingid mylogp "CC(C)CCO"
```

Result

```
id      logP
1       1,13
```

Without training the result is:

```
id      logP
```

```
1       1,09
```

Chemical Terms

Chemical Terms are available from Chemical Terms Evaluator or from Instant JChem. The *method* and *trainingid* parameters can be used in Chemical Terms Evaluator as well:

```
evaluate -e "logp('method:user trainingid:[library name]')" "[input
file/string]"
```

Example

```
evaluate -e "logp('method:user trainingid:mylogp')" "CC(C)CCO"
```

Instant JChem

You can also apply your log*P* training library via Chemical Terms in Instant JChem.

- Choose the 'New Chemical Terms Field icon' on the panel on the right side.
- Type the chemical term into the window, use the parameters method and trainingid. Do not forget to adjust the Name, the Type and the DB Column Name.

Example

The following figure presents the usage of log*P* training in the 'New Chemical terms' window. The expression

```
logP('method:user trainingid:mylogp')
```

defines that the plugin use the user defined log*P* training library *myplogp*.

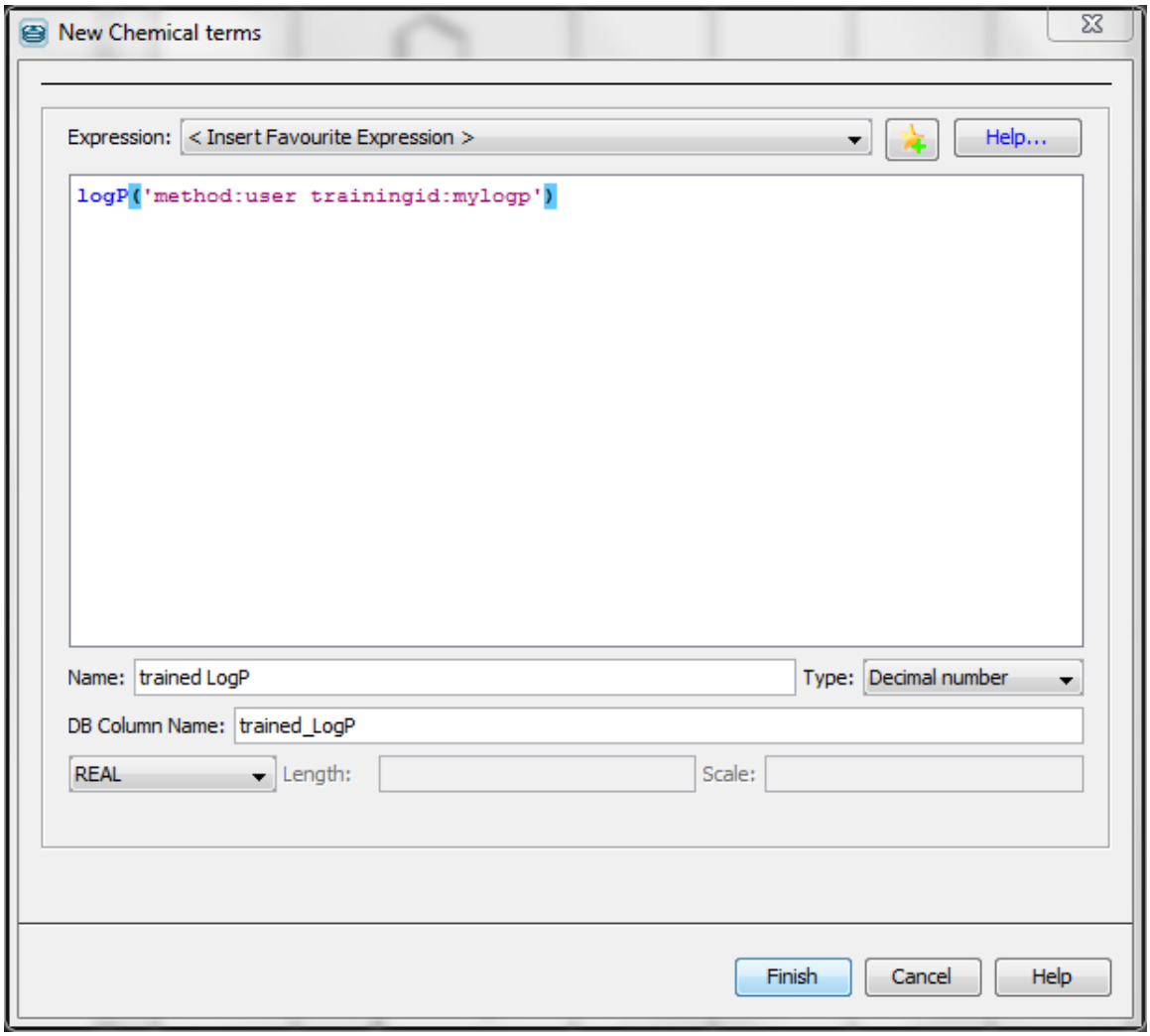


Fig. 3 Using Chemical Terms function for training in Instant JChem

Part of the results of this calculation is presented below. You can see the difference between the untrained (column *LogP*) and trained (column *trained LogP*) values.

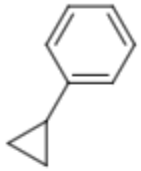
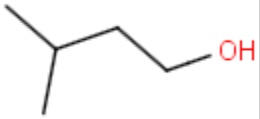
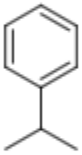
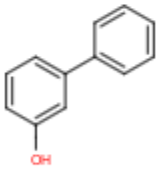
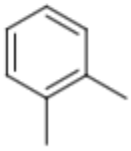
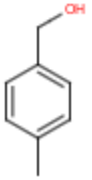
...	CdId	Structure	Mol Weight	Formula	LogP	trained LogP
5	5		118,18	C9H10	2,75	2,87
6	6		88,15	C5H12O	1,09	1,13
7	7		120,19	C9H12	3,22	3,34
8	8		170,21	C12H10O	3,32	3,31
9	9		106,17	C8H10	3,00	3,12
10	10		122,16	C8H10O	1,72	1,79

Fig. 4 JChem table showing the untrained and trained log*P* values