Manual of
SPCI (structural and physico-chemical interpretation) open-source software
version 1.0.0

Version (date)	Changes and comments
0.1.0 (02.02.2015)	Changes from alpha version: 1. More precise default SMARTS was added. 2. Cross-validation calculation was sped up and intermediate predictions are saved in text files. 3. Compounds, which cause errors in calculation of atomic properties with Chemaxon cxcalc tool, are excluded from further modeling. 4. Intermediate results of fragments contributions calculation are saved to a text file.
0.1.1 (07.02.2015)	1. Fixed errors in text output of intermediate results of fragments contributions calculation. 2. Fixed error in loading of file with descriptors on 32-bit platforms.
0.1.2 (13.05.2015)	Fixed error in loading of a file with descriptors on 32-bit platforms.
0.1.3 (22.07.2015)	Added two automatic fragmentation schemes: detection of i) all rings and ii) Murcko frameworks.
0.1.4 (21.01.2016)	1. Added automatic fragmentation scheme based on SMARTS to cleave bonds. 2. Changed routine of calculation of simplexes with hydrogen bonding labels. Simplexes with no labels A (acceptor), D (donor) or AD (acceptor and donor) are discarded. 3. Added support of multiple modeling properties in a single sdf-file (if property is missing put NA and such compounds will be discarded during modeling). 4. Default format of descriptors files is svm now (more affective to store sparse data). 5. GUI changes: - 10 last paths of loaded sdf-files are stored and can be easily loaded again - user may specify the number of cores to use for model building (default value is max_cores - 1) Note. Backward compatibility is broken, old projects cannot be opened with this new version due to changes in project structure.
0.1.5 (09.05.2016)	1. Predict module (the tab “Predict”) was added to predict properties of external datasets with the built models. 2. Project structure was changed: intermediate files required for calculation of non-user contributions are stored in the main project dir. 3. SiRMS submodule was updated. 4. Verbosity of output messages was decreased. 5. Many bug fixes (e.g. fixed calculation of specificity in classification models which crash the program).
1.0.0 (02.07.2018)	1. RDKit is used as a backend instead of Indigo for molecule fragmentation. 2. The multiple undersampling approach was implemented to model imbalanced data sets. 3. Default SiRMS descriptors were changed to min_atoms = 2 and max_atoms = 6. 4. SiRMS descriptors were updated to the version 1.1.2. 5. SVR makes grid search over epsilon parameter. 6. Added `utils` dir which contains useful scripts not used by GUI. 7. Many small fixes and improvements.

Overview 3
  Concept and citations 3
  Structural (doesn’t require Chemaxon) and physicochemical interpretation (Chemaxon required) 3
Installation and launch 3
Workflow 5
  Step 0. Data preparation and project start 5
  Step 1. Build models. 5
  Step 2. Calculation of the fragment contributions 6
  Step 3. Plot contributions and save plot in png file 6
  Step 3 (alternative visualization with R) 7
  Step 4. Predict properties of new compounds with built QSAR model 8

Overview

The SPCI software was designed for (semi)automatic extraction of structural features and their contributions to an investigated property from chemical datasets. It's a bunch of Python3 scripts. In order to simplify the usage a GUI was developed. It has a limited number of options but suits well for most needs.

Concept and citations

The idea of structural and physico-chemical interpretation along with comparison with MMP approach is disclosed in the following references. Please cite them if you use this software.

1) Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. Molecular Informatics **2013**, 32, 843-853.

2) Polishchuk, P., Tinkov, O., Khristova, T., Ognichenko, L., Kosinskaya, A., Varnek, A., Kuz'min, V., Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. Journal of Chemical Information and Modeling, **2016**, 56(8), 1455-1469.

Structural (doesn't require Chemaxon) and physicochemical interpretation (Chemaxon required)

Structural interpretation returns only overall fragments contributions while physicochemical interpretation can additionally estimate contributions of some physico-chemical factors (electrostatic, hydrophobic, hydrogen bonding and dispersive terms). Installed Chemaxon is required for physico-chemical interpretation.

If you have a license for Chemaxon check whether it is installed and PATH variable is correctly configured (to launch standardize and cxcalc from command line). If standardize or cxcalc command are not recognized from your command line, add JChem bin folder to PATH variable.

Installation and launch

The SPCI program is running under Python 3.4 with several dependencies: rdkit (2017.09), matplotlib (2.0.2), numpy (1.13.3), scipy (0.19.1), scikit-learn (0.19.1).

Since the SPCI program uses RDKit it is recommended to install RDKit and all other dependencies under Anaconda environment (http://www.rdkit.org/docs/Install.html). Therefore all installation and running steps are identical on Windows and Linux platforms.

To install SPCI clone the repository, initialize and update submodules:

```
git clone https://github.com/DrrDom/spci

git submodule init

git submodule update
```

To update the existed repository run sequentially:

```
git pull origin master

git submodule update
```

To launch the program - initialize RDKit conda environment and run `spci.py` script:

```
source activate my-rdkit-env
```

```
python3 spci.py
```

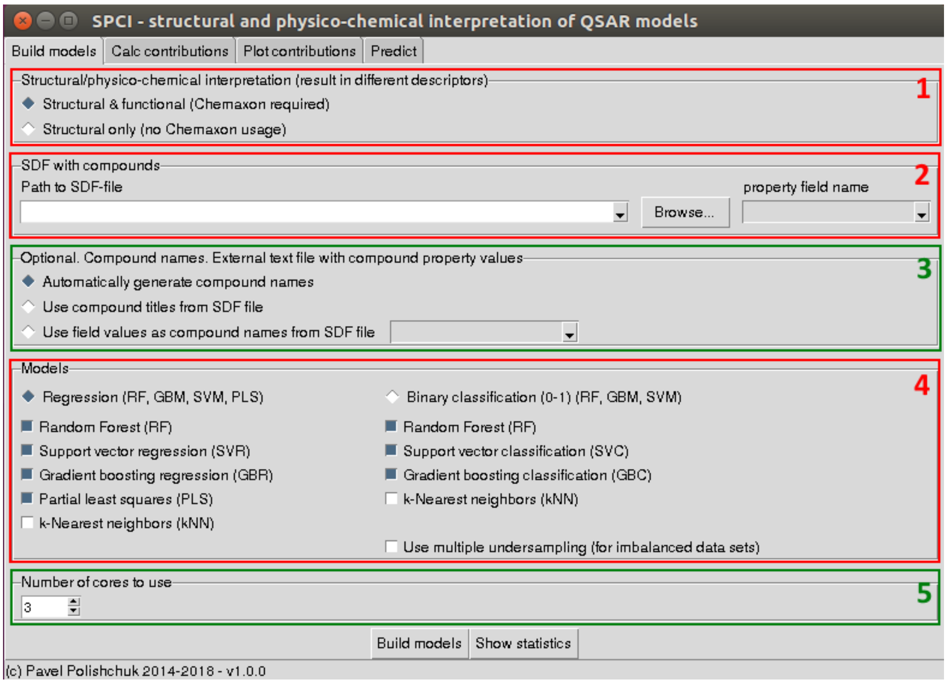
Workflow

Step 0. Data preparation and project start

For analysis a single sdf file is required, which contains compound structures and corresponding end-point values in a field of the sdf file. The input sdf file may contain several modeling properties. The structures should be checked and standardized. This is absolutely required in the case of only structural interpretation model, otherwise results can be distorted because standardization influence calculated descriptors. In the case of physicochemical interpretation will be chosen structures will be checked and standardized by Chemaxon utilities automatically.

Copy the sdf file in a separate dir which will be the project dir.

Step 1. Build models.



Structural/physico-chemical interpretation (result in different descriptors)
● Structural & functional (Chemaxon required)
○ Structural only (no Chemaxon usage)

SDF with compounds
Path to SDF-file
property field name

Optional. Compound names. External text file with compound property values
● Automatically generate compound names
○ Use compound titles from SDF file
○ Use field values as compound names from SDF file

Models
● Regression (RF, GBM, SVM, PLS)
☑ Random Forest (RF)
☑ Support vector regression (SVR)
☑ Gradient boosting regression (GBR)
☑ Partial least squares (PLS)
☐ k-Nearest neighbors (kNN)
○ Binary classification (0-1) (RF, GBM, SVM)
☑ Random Forest (RF)
☑ Support vector classification (SVC)
☑ Gradient boosting classification (GBC)
☐ k-Nearest neighbors (kNN)
☐ Use multiple undersampling (for imbalanced data sets)

Number of cores to use
3

(c) Pavel Polishchuk 2014-2018 - v1.0.0

*Red fields are required, green ones are optional.

1. Choose the type of analysis to perform.

2. Specify path to sdf file with compounds and choose the name of the field containing property values. Paths to 10 last opened sdf files are stored in the dropdown list.

3. Optional. It is required if you want to save original compounds names.

4. Choose the desired type of models. All listed models will be build with optimal parameters selected by the grid search in 5-fold cross validation procedure. For the imbalanced classification data sets one may chose option to build multiple models using undersampling. For large and very imbalanced data sets this can substantially increase computational time and model size on a disk.

5. Optional. The number of cores to be used during models building. By default all cores minus 1 will be used.

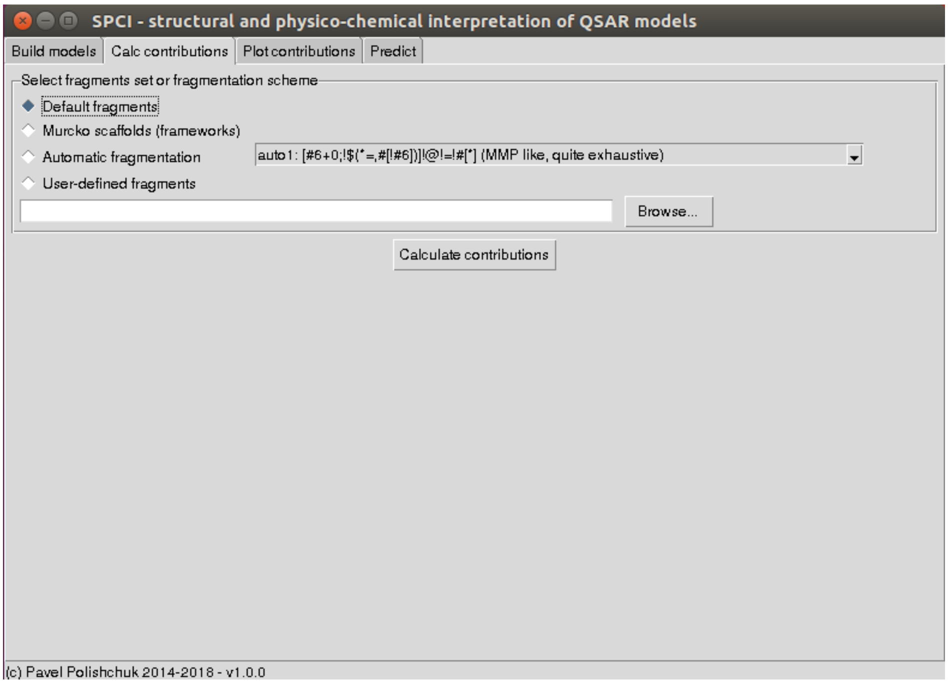
IMPORTANT:

To open an earlier created project you should choose the input sdf file and property field name.

Step 2. Calculation of the fragment contributions

Contributions will be calculated only for those models specified on the previous tab.

If models had reliable predictive ability estimated by 5-fold cross-validation (Show statistics button on the first tab) one may calculate fragments contributions.


SPCI - structural and physico-chemical interpretation of QSAR models
Build models
Calc contributions
Plot contributions
Predict
Select fragments set or fragmentation scheme
Browse...
Calculate contributions
(c) Pavel Polishchuk 2014-2018 - v1.0.0

● Default fragments
○ Murcko scaffolds (frameworks)
○ Automatic fragmentation auto1: [#6+0;!$(*=,#[!#6])]!@!=!#[*] (MMP like, quite exhaustive)
○ User-defined fragments

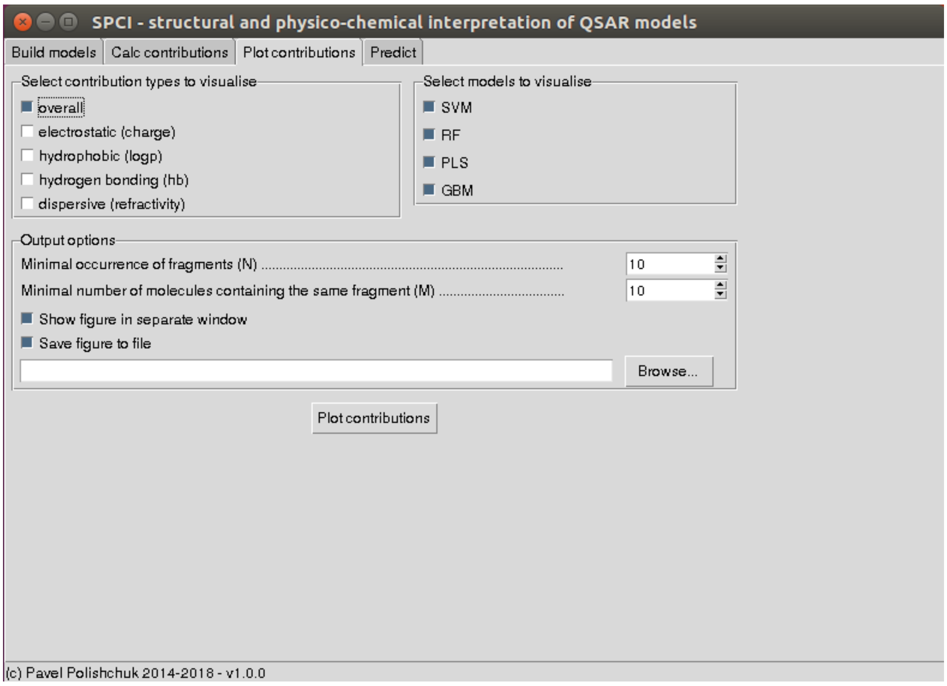
Default fragments – common functional groups and rings (specified in default.smarts file in program folder);

Murcko frameworks – automatically detect Murcko frameworks;

automatic fragmentation – split molecules on maximum three parts on bonds match SMARTS: [#6+0;!$(*=,#[!#6])]!@!=!#[*];

user-defined fragments represented in SMILES or SMARTS notation, look at default.smarts as an example of file format (Note: SDF input was not tested).

Step 3. Plot contributions and save plot in png file


SPCI - structural and physico-chemical interpretation of QSAR models
Build models
Calc contributions
Plot contributions
Predict
Select contribution types to visualise
Select models to visualise
Output options
Browse...
Plot contributions
(c) Pavel Polishchuk 2014-2018 - v1.0.0

☑ overall
☐ electrostatic (charge)
☐ hydrophobic (logp)
☐ hydrogen bonding (hb)
☐ dispersive (refractivity)

☑ SVM
☑ RF
☑ PLS
☑ GBM

Minimal occurrence of fragments (N) 10
Minimal number of molecules containing the same fragment (M) 10
☑ Show figure in separate window
☑ Save figure to file

After the project was opened all available models will be listed in order to choose them for visualization of fragments selected on the second tab.

Overall contributions (structural interpretation). Other types of contributions are used for physicochemical interpretation (to show the contribution of separate physicochemical factors).

Step 3 (alternative visualization with R)

You may create alternative visualization in order to customize plot output:

1) by yourself (parsing *_frag_contribution.txt files with calculated contributions located inside the modeling property folders and applying any of available tools);

2) by using the developed web-base tool which is suitable for visualization of relatively small set of fragments (number of columns in *_frag_contribution.txt should be less than 100k) otherwise it will be very slow or even impossible;

Full version - http://158.194.101.252:3838/spci-vis/

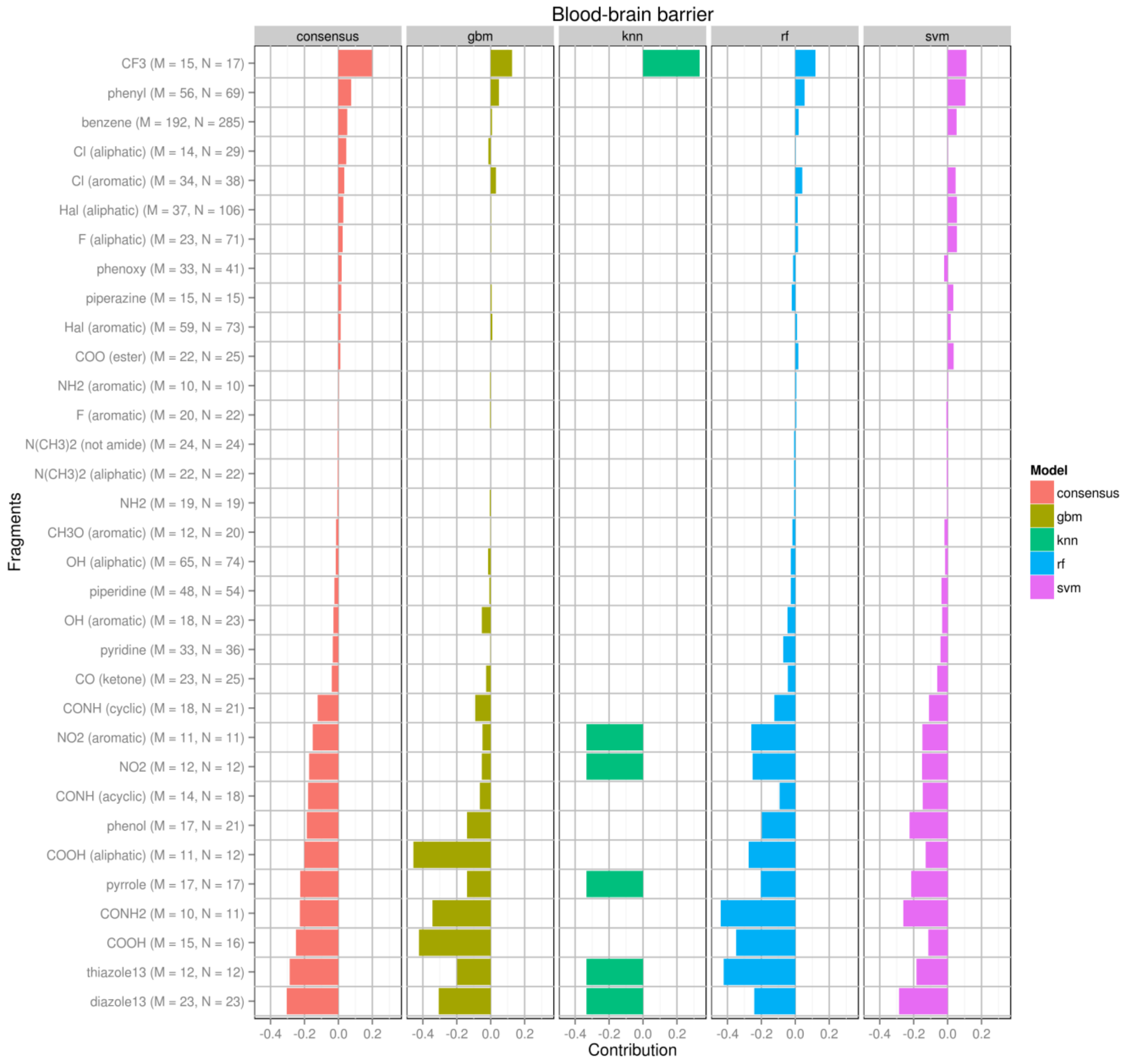
Demo version to just play with pre-loaded data - http://158.194.101.252:3838/spci-vis-demo/

3) by using `rspci` R package on your local machine. This package contains functions to facilitate data reading, modification, filtering and drawing. It has more options that the web-based tool and recommended for advanced graphics to R users.

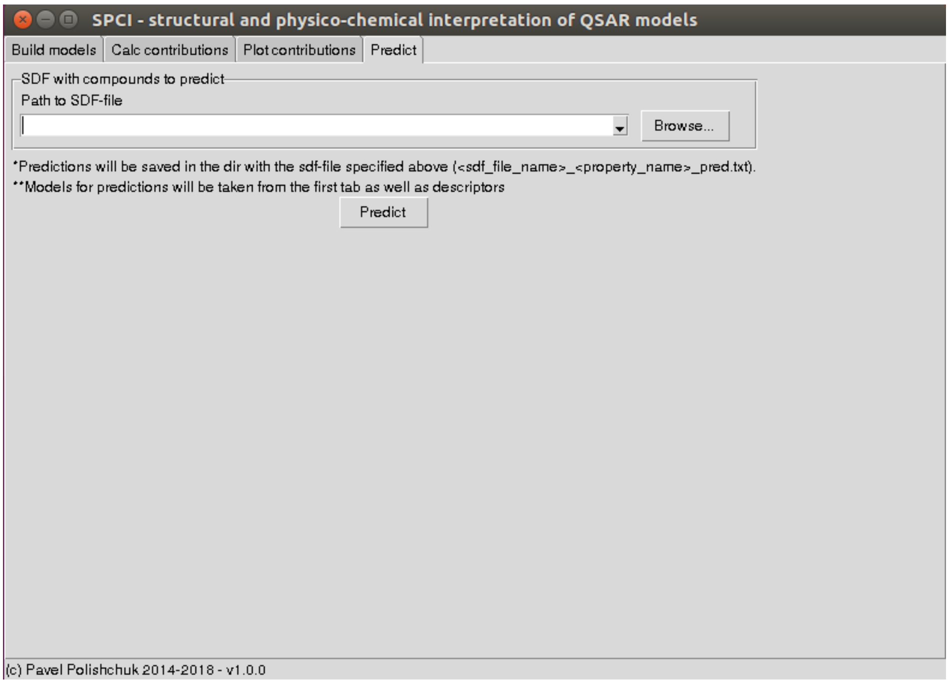
To install `rspci` package call from R console:

```
devtools::install_github(“DrrDom/rspci”)
```

Below is an example of a contribution plot:



Step 4. Predict properties of new compounds with built QSAR model


SPCI - structural and physico-chemical interpretation of QSAR models
Build models
Calc contributions
Plot contributions
Predict
SDF with compounds to predict
Path to SDF-file
Browse...
Predict
(c) Pavel Polishchuk 2014-2018 - v1.0.0

*Predictions will be saved in the dir with the sdf-file specified above (<sdf_file_name>_<property_name>_pred.txt).

**Models for predictions will be taken from the first tab as well as descriptors

Specify the path to the sdf-file with an external set of compounds and choose models and the end-point for prediction from the first tab.

Applicability domain is estimated for each compound based on bounding box – if at least one descriptor has a value outside the min-max range of this descriptor values for the training set compounds the compound is out of the domain applicability.