Date: 2024/04/30
Yuichi Katsumura
SAS Institute Japan Ltd.

How to use Toolkit for Materials Informatics add-in

Note: This document is translated using Microsoft Translator.

1. What is the purpose of this add-in?

This add-in for materials informatics and chemical informatics.
Here's what you can do:

- Generate molecular descriptors
- Generate a fingerprint
- Draw molecules
- Preprocess descriptors for use in predictive models

Generating molecular descriptors, generating fingerprints, drawing molecules, etc. can be done in RDKit.

2. System Requirements

- JMP 18 or later
- Available in English and Japanese only
- Windows OS
- The following Python packages must be installed.
- RDKit
- Numpy
- Pandas

This add-in uses Python integration (new in JMP 18).

3. Test environment

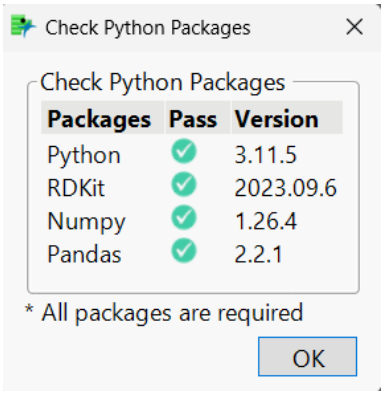
- JMP 18.0
- Windows 11
- English, Japanese
- Python 3.11.5
- RDKit 2023.09.6

4. Installing Python packages

1) Installing RDKit
   1. Select File > New > JSL Script
   2. Enter the script below and select Edit > Run Script

```
Python Install Packages("numpy pandas rdkit");
```

2) Installing the add-in
   1. In File > Open, select Toolkit for Materials Informatics.jmpaddin and install it.
   2. In the Add-Ins menu, select Toolkit for Materials Informatics > Check Python Packages. If you see a window like the one below, there is no problem.

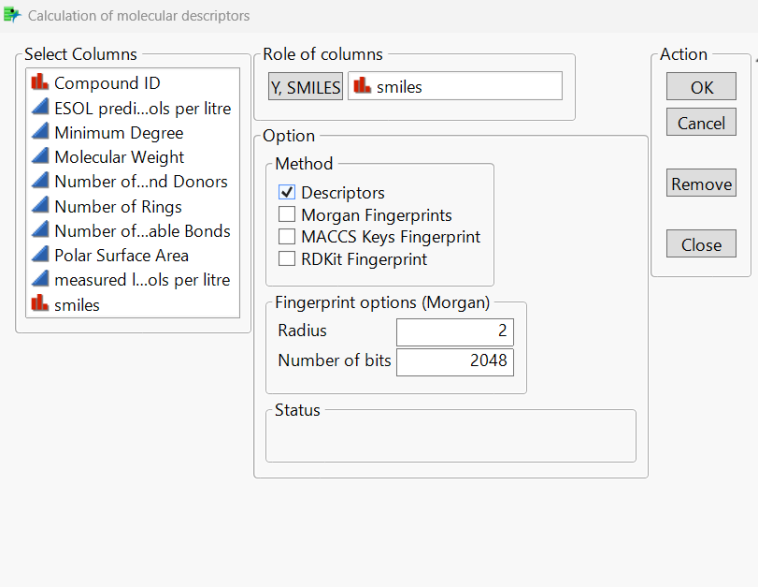


5. How to use

In this documentation, theESOL_delaney-processed.csv (Lambard) is used.

[1] Molecular Descriptor/Fingerprint Generation
Generate molecular descriptors/fingerprints from SMILES.

1) Select Add-ins > Toolkit for Materials Informatics > Descriptor Calculation.
2) Specify the column in which SMILES is entered as "Y,SMILES". Check the descriptors and fingerprints you want to generate from the list of method selections.



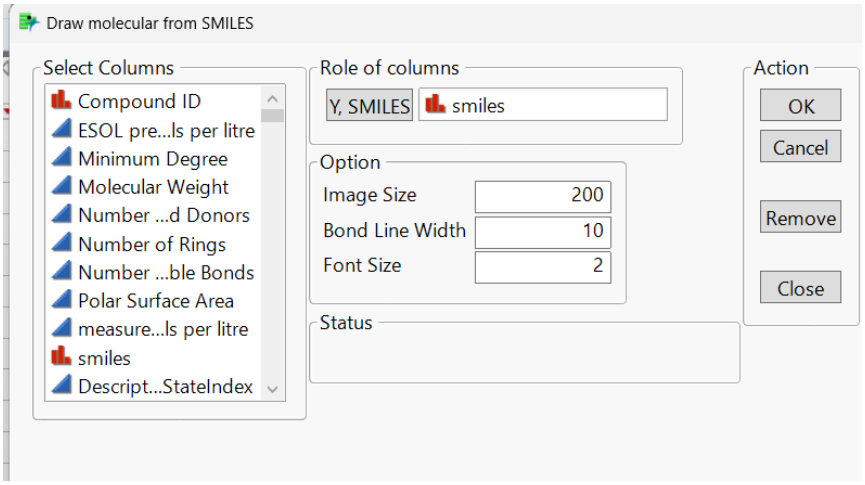
- Descriptors and fingerprints are generated using RDKit.
  https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors
- If you select Descriptors, the descriptor for the "List of Available Descriptors" section of the above page will be generated.
- For Morgan Fingerprints、MACCS Keys Fingerprint、RDKit Fingerprint, please refer to "List of Available Fingerprints".
- For Morgan Fingerprints, specify the radius and the number of bits.

3) Click the OK button to generate the selected features.
4) When you're done, click the "Close" button.

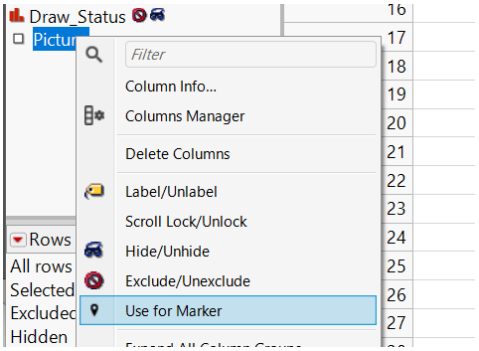
[2] Drawing molecules

Draw molecules from SMILES and add them to the data table.

1) Select Add-ins > Toolkit for Materials Informatics > Drawing Molecules.
2) Specify the column in which SMILES is entered as "Y,SMILES". Specifies the size of the image to be generated.
3) When you click the OK button, a column with the generated image will be added.



4) RDKit is used to generate images.
https://www.rdkit.org/docs/source/rdkit.Chem.Draw.html#module-rdkit.Chem.Draw
5) You can use an image as a marker by right-clicking on a column in the list of column names and selecting Use For Marker.



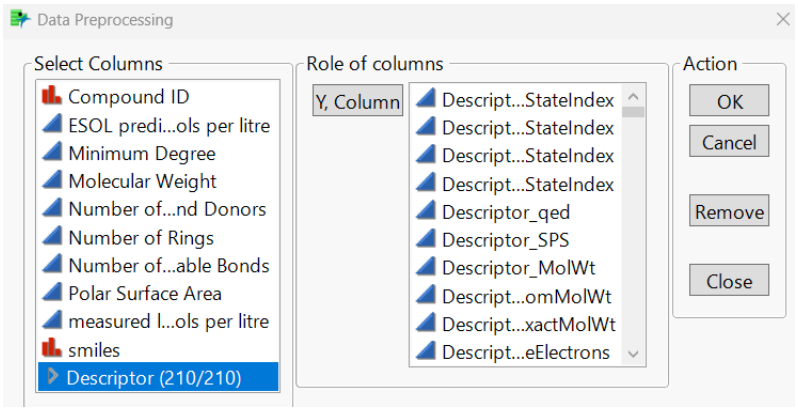
6) If you're using an image as a marker and the image is hard to see, try increasing the size of the marker or adjusting the line thickness when generating the image.

[3] preprocessing

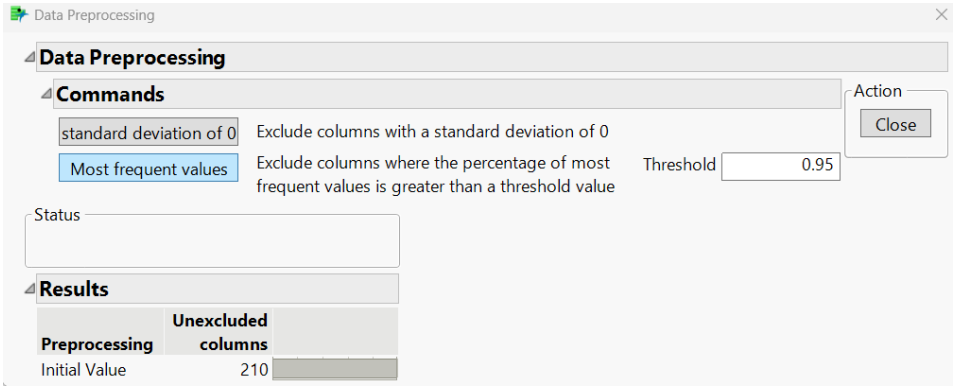
You can perform the following preprocessing:

- Exclude columns with 0 standard deviation
- The proportion of the most frequent values is above the threshold

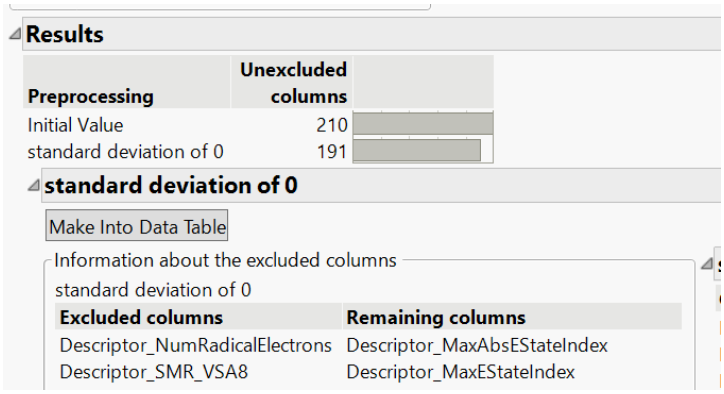
1) Select Add-ins > Toolkit for Materials Informatics > Preprocessing.
2) Specify the column in which the descriptor, fingerprint, etc. are entered as "Y, column". Click the OK button.



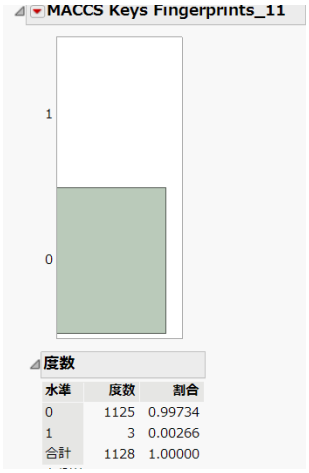
3) If you click the button with standard deviation of 0 and most frequent values, the calculation will be performed and the column will be excluded.



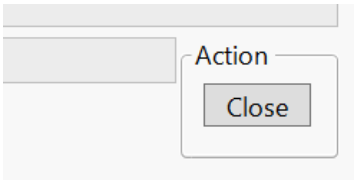
4) For example, if you click the standard deviation of 0 button, the standard deviation is calculated for each column. Columns with a standard deviation of 0 are excluded. The columns that were excluded are shown in the Results outline below the report.



5) If you click Most Frequent Values, for example, if a column contains values of 0 and 1, and the frequency of 0 is greater than or equal to the value set for Threshold, such as 99% (the default is 95%), the column is excluded.



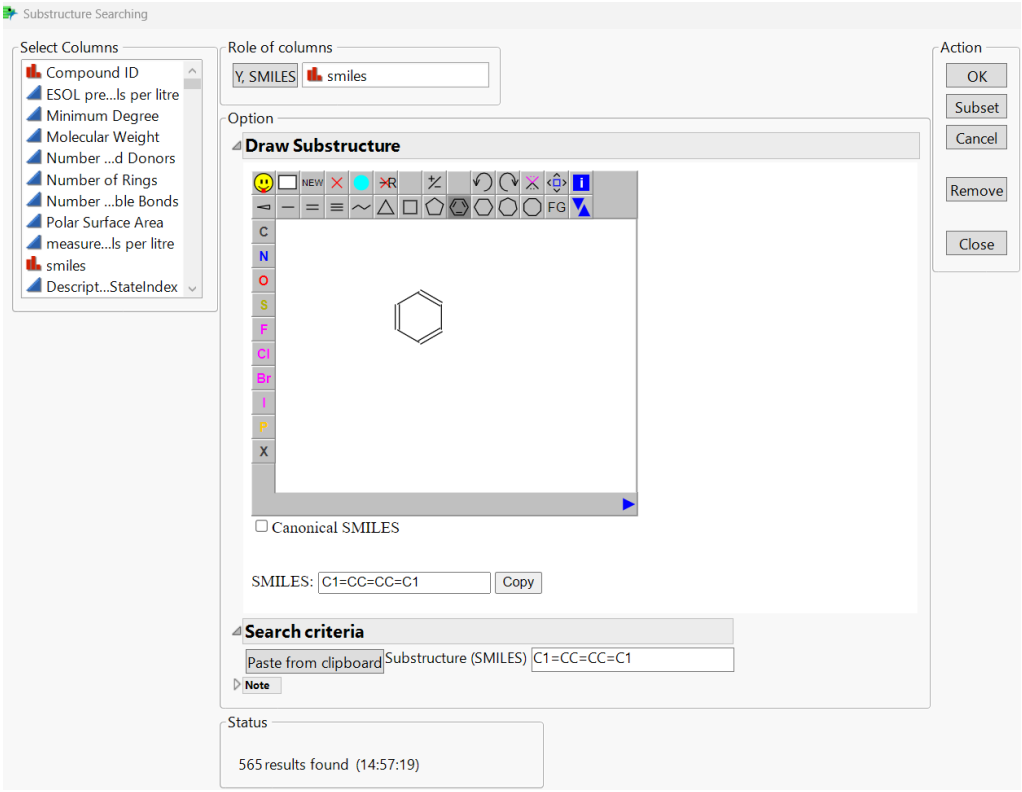
Click the Close button to close the window.



[4] Searching for Substructures

You can search and subset the data by whether the specified SMILES are included in the substructure or not.

1) Select Add-ins > Toolkit for Materials Informatics > Substructure Searching.
2) Specify the column in which SMILES is entered as "Y,SMILES" and draw the structure you want to find in Drawing Substructures. When SMILES is displayed, click the "Copy" button and click the "Paste from clipboard" button.
3) When you click the OK button, a column called Substructure: <SMILES>" is added to the data table, and the rows with the value True are the substructure-matched rows. Click the Subset button to subset only the matched rows.



The drawing of the substructure uses the following software:

- JSME Molecule Editor - 3-Clause BSD License

Please refer to the following page for how to use it.

WE ARE HELP (jsme-editor.github.io)

References

RDKit: Open-Source Cheminformatics Software. [Online] [Cited: 9 7, 2023.] https://rdkit.org/.

Lambard, Guillaume. [Online] [Cited: 9 7, 2023.] https://github.com/GLambard/Molecules_Dataset_Collection.