Prediction of SE (side effect) or TI (therapeutic) based on Edit-Distance between paths, which is extracted from KEGG pathway maps.

# Overview

These folders are used to make SE/TI predictions based on Pathway similarity. Pathway similarity is calculated using a metric called Edit-Distance.

# Description

There are 24 folders, which can be divided into the following two parts;

1. Training prediction models.

2. Predicting SE/TI actually.

1. Training prediction models.

Folders with leading numbers of 0-11 are for model training and basically contain data files and code files in ipynb format, with a description of each ipynb file at the top (ex. the description of using columns of input files).

The part of training prediction models can be further divided into three parts;

1. Obtaining datasets from KEGG, SIDER and DrugBank.
2. Processing the data (explanatory variables and response variable).

b-1) Extracting paths from KEGG pathway maps.

b-2) Calculating Edit-Distance between paths in order to process the data of explanatory variables.

b-3) Associating paths, extracting from KEGG pathway maps, with SE/TI by using the dataset from SIDER and DrugBank and binarizing it in order to process the data of response variable.

b-4) Selecting features for explanatory variables (and performing PCA).

b-5) Splitting into training and test data.

1. Training models, using the processed data.

Feature selection is performed in two ways (hereafter referred to as Method A and Method B), the procedure for running the code (Method A and B) is as follows.

* Method A

[0\_KEGG\_Pathway\_xmlfile] (In phase **a**)

([0\_1\_scraping\_pathway\_link.ipynb]->[0\_2\_scraping\_pathway\_xmlfile.ipynb])

=>

[1\_Paths\_from\_KEGG\_Pathway\_code] (In phase **b-1**)

([1\_1\_xml\_analysis.ipynb]->[1\_2\_paths\_extraction.ipynb]->[1\_3\_path\_id.ipynb])

=>

[3\_Calc\_Edit\_Distance] (In phase **b-2**)

([3\_1\_Calc\_Edit\_Distance.ipynb])

=>

[8\_Integration\_SE\_TI\_Target\_code] (In phase **b-3**)

([8\_binary\_SE\_TI\_Target.ipynb])

=>

[4\_Feature\_extraction] (In phase **b-4**, **5**)

([4\_0\_train\_test\_count\_TI.ipynb]->[4\_1\_Feature\_extraction\_TI.ipynb]-> [4\_1\_Feature\_extraction\_SE.ipynb]-> [4\_0\_train\_test\_count\_SE.ipynb])

=>

[4\_Feature\_extraction\_PCA] (In phase **b-4**)

([4\_2\_PCA\_TI.ipynb]->[4\_2\_PCA \_SE.ipynb])

=>

[10\_build\_LGBM\_code\_PCA] (In phase **c**)

([10\_1\_build\_LGBM\_TI.ipynb]->[10\_1\_build\_LGBM\_SE.ipynb])

* Method B

[0\_KEGG\_Pathway\_xmlfile] (In phase **a**)

([0\_1\_scraping\_pathway\_link.ipynb]->[0\_2\_scraping\_pathway\_xmlfile.ipynb])

=>

[1\_Paths\_from\_KEGG\_Pathway\_code] (In phase **b-1**)

([1\_1\_xml\_analysis.ipynb]->[1\_2\_paths\_extraction.ipynb]->[1\_3\_path\_id.ipynb])

=>

[3\_Calc\_Edit\_Distance] (In phase **b-2**)

([3\_1\_Calc\_Edit\_Distance.ipynb])

=>

[8\_Integration\_SE\_TI\_Target\_code] (In phase **b-3**)

([8\_binary\_SE\_TI\_Target.ipynb])

=>

[4\_Feature\_extraction] (In phase **b-4**, **5**)

([4\_0\_train\_test\_count\_TI.ipynb]->[4\_1\_Feature\_extraction\_TI.ipynb]-> [4\_1\_Feature\_extraction\_SE.ipynb]-> [4\_0\_train\_test\_count\_SE.ipynb])

=>

[4\_Feature\_extraction\_nonPCA] (In phase **b-4**)

([Target\_sub\_cluster.ipynb]->[4\_2\_Clustering\_TI.ipynb]->[4\_2\_Clustering\_SE.ipynb])

=>

[10\_build\_LGBM\_code\_nonPCA] (In phase **c**)

([10\_1\_build\_LGBM\_TI.ipynb]->[10\_1\_build\_LGBM\_SE.ipynb])

2. Predicting SE/TI actually.

Folders with leading numbers of 12-14 are for prediction, containing code files in Python and ipynb format and data files.

These folders were analyzed for a gene (has:2913, GRM3) as an example. The analysis was divided into the following three stages.

1. Getting input components.
2. Predicting SE/TI for components.
3. Analyzing the result.

In this part, Main folders are [13\_Prediction\_SE\_TI\_code\_PCA] and [13\_Prediction\_SE\_TI\_code\_nonPCA], which contain the code ([KEGG\_Based\_TI\_SE\_Predict\_Model\_functions\_Method\_A.py]

and [KEGG\_Based\_TI\_SE\_Predict\_Model\_functions\_Method\_B.py]) with self-made functions written in.

In [KEGG\_Based\_TI\_SE\_Predict\_Model\_functions\_Method\_A.py], we use the following 14 files.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_edge\_DrugBank.csv]

In this notebook, we used this file with the exception of the column 'name'.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_node\_DrugBank.csv]

In this notebook, we used this file with the exception of the column ‘arrow’.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_node\_DrugBank\_group.csv]

In this notebook, we used all columns in this file.

* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_binary\_TI.npz]
* [6\_PCA\_model/X\_PCA\_model\_TI\_\*.pkl]
* [11\_LGBM\_TI\_PCA/model\_ti\_\*.pkl]
* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_ID\_name\_TI.csv]

In this notebook, we used all columns in this file.

* [4\_Feature\_extraction/output/Train\_Test\_count\_TI.csv]

In this notebook, we only used the column ‘ID’ and 'use\_ID' in this file.

* [3\_Calc\_Edit\_Distance/output/KEGG\_ID\_index.csv]

In this notebook, we used this file with the exception of the column ‘has\_map’.

* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_binary\_SE.npz]
* [6\_PCA\_model/X\_PCA\_model\_SE\_\*.pkl]
* [11\_LGBM\_SE\_PCA/model\_se\_\*.pkl]
* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_ID\_name\_SE.csv]

In this notebook, we used all columns in this file.

* [4\_Feature\_extraction/output/Train\_Test\_count\_SE.csv]

In this notebook, we only used the column ‘ID’ and 'test' in this file.

In [KEGG\_Based\_TI\_SE\_Predict\_Model\_functions\_Method\_B.py], we use the following 12 files.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_edge\_DrugBank.csv]

In this notebook, we used this file with the exception of the column 'name'.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_node\_DrugBank.csv]

In this notebook, we used this file with the exception of the column ‘arrow’.

* [1\_Paths\_from\_KEGG\_Pathway\_code/output/all\_target\_node\_DrugBank\_group.csv]

In this notebook, we used all columns in this file.

* [4\_Feature\_extraction\_nonPCA/output/TI\_Feature\_Cluster\_ID.csv]

In this notebook, we only used the column ‘ID’ and ‘SE\_ID’ in this file.

* [11\_LGBM\_TI\_nonPCA/model\_ti\_\*.pkl]
* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_ID\_name\_TI.csv]

In this notebook, we used all columns in this file.

* [4\_Feature\_extraction/output/Train\_Test\_count\_TI.csv]

In this notebook, we only used the column ‘ID’ and 'use\_ID' in this file.

* [3\_Calc\_Edit\_Distance/output/KEGG\_ID\_index.csv]

In this notebook, we used this file with the exception of the column ‘has\_map’.

* [4\_Feature\_extraction\_nonPCA/output/SE\_Feature\_Cluster\_ID.csv]

In this notebook, we only used the column ‘ID’ and ‘SE\_ID’ in this file.

* [11\_LGBM\_SE\_nonPCA/model\_se\_\*.pkl]
* [9\_Integration\_SE\_TI\_Target\_datafile/Y\_ID\_name\_SE.csv]

In this notebook, we used all columns in this file.

* [4\_Feature\_extraction/output/Train\_Test\_count\_SE.csv]

In this notebook, we only used the column ‘ID’ and 'test' in this file.

Next, I will explain the role of folders which saves datafiles.

- [2\_Paths\_from\_KEGG\_Pathway\_datafile]

This contains a file with output from running the code

[1\_Paths\_from\_KEGG\_Pathway\_code].

- [5\_X\_train\_test\_datafile]

This contains a file with output from running the code [4\_Feature\_extraction].

- [5\_X\_train\_test\_datafile\_PCA]

This contains a file with output from running the code [4\_Feature\_extraction\_PCA].

- [5\_X\_train\_test\_datafile\_nonPCA]

This contains a file with output from running the code [4\_Feature\_extraction\_nonPCA].

- [6\_PCA model]

This contains a file with output from running the code [4\_Feature\_extraction\_PCA].

- [7\_Target\_Drug\_SE\_TI\_from\_SIDER\_DrugBank]

This contains two files. One is [tbl\_drug\_target\_TI\_SE.pkl], a formatted file of data downloaded from SIDER and DrugBank and the other is [Use\_SE\_177.csv], a file of SE list which is used for prediction.

- [9\_Integration\_SE\_TI\_Target\_datafile]

This contains a file with output from running the code [8\_Integration\_SE\_TI\_Target\_code].

- [11\_LGBM\_(SE/TI)\_PCA]

This contains a file with output from running the code [10\_build\_LGBM\_code\_PCA].

- [11\_LGBM\_(SE/TI)\_nonPCA]

This contains a file with output from running the code [10\_build\_LGBM\_code\_nonPCA].

- [12\_input\_to\_LGBM]

This contains an example of a data file to be input to [13\_Prediction\_SE\_TI\_code\_(non)PCA].

In this case, as an example, we made a prediction of SEs and TIs for GRM3 (KEGG ID : 2913).

- [14\_output\_from\_LGBM]

This contains a file with output as a result from running the code

[13\_Prediction\_SE\_TI\_code\_(non)PCA].

# Requirement

Python 3.8.8