# System requirements

The notebooks have been tested on the following systems:

* Linux: Ubuntu 18.04
* Microsoft windows 10

## Dependencies

Users should install the following packages prior to running the notebooks

matplotlib

numpy

pandas

scipy

seaborn

sinfo

sklearn

tensorflow

shap

umap-learn

# Installation guide

Download *chemPredPD.tar.gz*

*tar -zxvf chemPredPD.tar.gz*

To install the neurocrick module that contains the utilities functions used in the project

*cd <path\_to\_directory\_holding\_classifier>\chemPredPD\neurocrick*

*pip install -e .*

This takes approximately under 5 minutes to install.

# Demo

Instructions to run on data.

*cd <path\_to\_directory\_holding\_classifier>\chemPredPD*

The notebooks have been set up to run with the demo data in *chemPredPD\demo\_data*

The output of the runs with the demo data will be saved under *chemPredPD\demo\_output*

For the demo, open the notebook in jupyter notebook and run all cells.

|  |  |  |
| --- | --- | --- |
| **Notebook** | **Description** | **Demo output saved to folders under** demo\_output |
| *tabular\_model\_5way\_skfold.ipynb* | Running Stratified K-Fold Cross Validator : predicting five classes; four disease and one healthy state, using tabular data based on the nucleus, mitochondria and lysosome features | tab\_data\_5way\_skfold |
| *tabular\_model\_5way\_shap.ipynb* | Predicting five classes; four disease and one healthy state, using tabular data based on the nucleus, mitochondria and lysosomal features. Running shapley for the model saved | tab\_data\_5way\_SHAP |
| *coloc\_model\_5way.ipynb* | Predicting Control vs Oligomycin vs 3xSNCA vs oligomer vs rotenone using tabular data from mitochondria and lysosome colocalization | coloc\_tab\_data\_classify5way |
| *coloc\_model\_aggr\_mito. ipynb* | Predicting disease - Aggregation vs Mitochondrial toxicity using tabular data from mitochondria and lysosome colocalization | coloc\_tab\_data\_aggr\_mt |
| *coloc\_model\_aggr\_mito\_run\_shapley.ipynb* | Running shapley for the model saved by  coloc\_model\_aggr\_mito. ipynb | coloc\_tab\_data\_aggr\_mt |
| *coloc\_model\_aggr\_mito\_skfold.ipynb* | Running Stratified K-Fold Cross Validator : Predicting diseases - Agregation vs Mitochondrial toxicity using tabular data from mitochondria and lysosome colocalization | coloc\_aggr\_mt\_skfold |
| *image\_model\_5way\_shap.ipynb* | Predicting five classes; four disease and one healthy state, using segmented raw images of the nucleus, mitochondria and lysosome. Running shapley for the model saved | image\_data\_5way\_SHAP |
| *image\_model\_5way\_skfold.ipynb* | Running Stratified K-Fold Cross Validator : predicting five classes; four disease and one healthy state, using segmented raw images of the nucleus, mitochondria and lysosome | image\_data\_5way\_skfold |
| *image\_model\_genetic\_classifier.ipynb* | Predicting three classes; two disease (SNCAx3 versus PINK1) and one healthy state, using segmented raw images of the nucleus, mitochondria and lysosome. Running shapley for the model saved | image\_data\_genetic\_SHAP |
| *image\_model\_genetic\_classifier\_skfold.ipynb* | Running Stratified K-Fold Cross Validator : predicting three classes; two disease (SNCAx3 versus PINK1) and one healthy state, using segmented raw images of the nucleus, mitochondria and lysosome | Image\_data\_genetic\_SK\_fold |

Please find below under ‘instructions for use’ the expected output for each of the notebooks.

**Expected run time for demo on a "normal" desktop computer**

The tabular data classifiers take approximately 3.5 hours to run on a Microsoft Windows 10 system.

The 5 class image classifiers with the demo dataset required 8vCPU, 32GB memory, 8GB GPU memory. Using this configuration image\_model\_5way\_shap.ipynb has a run time of 2 hours and image\_model\_5way\_skfold.ipynb has a run time of 16.5 hours.

# Instructions for use

In the notebooks below, for **input** : Under the ‘Set paths’ section, update the variables

'data\_file' with the input file (including path) and

'path\_res' with the output folder path

**Notebook - tabular\_model\_5way\_shap.ipynb**

Predicting Control vs Oligomycin vs 3xSNCA vs oligomer vs rotenone using tabular data based on the nucleus, mitochondria and lysosomal features

Output:

The saved\_models folder contains the best model.

The plots folder contains png files generated with the following prefixes:

* CONFMATRIX\_count : confusion matrix (counts)
* CONFMATRIX\_fraction: confusion matrix (fraction)
* Model-perfor: plots showing the training and validation loss and accuracy

The plots folder also contains shapley summary plots for each class and shapley feature importance plot.

**Notebook - tabular\_model\_5way\_skfold.ipynb**

Running Stratified K-Fold Cross Validator : predicting five classes; four disease and one healthy state, using tabular data based on the nucleus, mitochondria and lysosome features

Output :

The output is saved in Plots and saved\_models folders created in the output directory specified.

In the plots folder :

- Model-perform.png : plot showing the training and validation loss and accuracy

The saved\_models folder contains the models saved from the Stratified K-Fold run and the accuracy scores.

**Notebook** **: coloc\_model\_5way.ipynb**

Predicting Control vs Oligomycin vs 3xSNCA vs oligomer vs rotenone using tabular data from mitochondria and lysosome colocalization.

Output :

Creates a ‘Plots’ and ‘saved\_models’ folder in the output directory specified.

The plots folder contains png files generated with the following prefixes :

- CONFMATRIX\_count : confusion matrix (counts)

- CONFMATRIX\_fraction : confusion matrix ( fraction)

- CONFMATRIX\_mixed-plates : both the above confusion matrices in a single plot.

- Model-perfor : plots showing the training and validation loss and accuracy

The saved\_models folder contains the best model.

.csv files containing the scaled training, validation and test datasets with the features are saved to the main output folder.

**Notebook** **: coloc\_model\_aggr\_mito.ipynb**

Predicting disease - Aggregation vs Mitochondrial toxicity using tabular data from mitochondria and lysosome colocalization

The output folder and files are in line with those generated by *coloc\_model5way.ipynb*

**Notebook - coloc\_model\_aggr\_mito\_run\_shapley. ipynb**

Running shapley for the model saved by *coloc\_model\_aggr\_mito. ipynb*

Input : Under set paths section, update the variable

'path\_res' with the results folder path set in coloc\_model\_aggr\_mito. ipynb

Output :

saves the shap values to the 'path\_res' as shap\_values\_\*.pkl

<path\_res>/plots contains the shapley summary plots for each class and shapley feature importance plot.

**Notebook - coloc\_model\_aggr\_mito\_skfold.ipynb**

Running Stratified K-Fold Cross Validator : Predicting diseases - Aggregation vs Mitochondrial toxicity using tabular data from mitochondria and lysosome colocalization

The saved\_models folder contains the models saved from the Stratified K-Fold run and the accuracy scores.

**Notebook - image\_model\_5way\_shap.ipynb**

Predicting control vs oligomycin vs 3xSNCA vs oligomer vs rotenone using segmented raw images of the nucleus, mitochondria and lysosome

Output:

The saved\_models folder contains the best model.

The plots folder contains png files generated with the following prefixes:

* CONFMATRIX\_count : confusion matrix (counts)
* CONFMATRIX\_fraction: confusion matrix (fraction)
* Model-perfor: plots showing the training and validation loss and accuracy

The plots folder also contains the shapley plot and the most representative images for each class

**Notebook - image\_model\_5way\_skfold.ipynb**

Predicting control vs oligomycin vs 3xSNCA vs oligomer vs rotenone using segmented raw images of the nucleus, mitochondria and lysosome

Output:

The saved\_models folder contains the models saved from the Stratified K-Fold run and the accuracy scores.

the plots folder contains the plot showing the training and validation loss and accuracy

**Notebook - image\_model\_genetic\_classifier.ipynb**

Predicting control vs 3xSNCA vs PINK1 using segmented raw images of the nucleus, mitochondria and lysosome

Output:

The saved\_models folder contains the best model.

The plots folder contains png files generated with the following prefixes:

* CONFMATRIX\_count : confusion matrix (counts)
* CONFMATRIX\_fraction: confusion matrix (fraction)
* Model-perfor: plots showing the training and validation loss and accuracy

The plots folder also contains the shapley plot and the most representative images for each class

**Notebook - image\_model\_genetic\_classifier\_skfold.ipynb**

Predicting control vs 3xSNCA vs PINK1 using segmented raw images of the nucleus, mitochondria and lysosome

Output:

The saved\_models folder contains the models saved from the Stratified K-Fold run and the accuracy scores.

the plots folder contains the plot showing the training and validation loss and accuracy