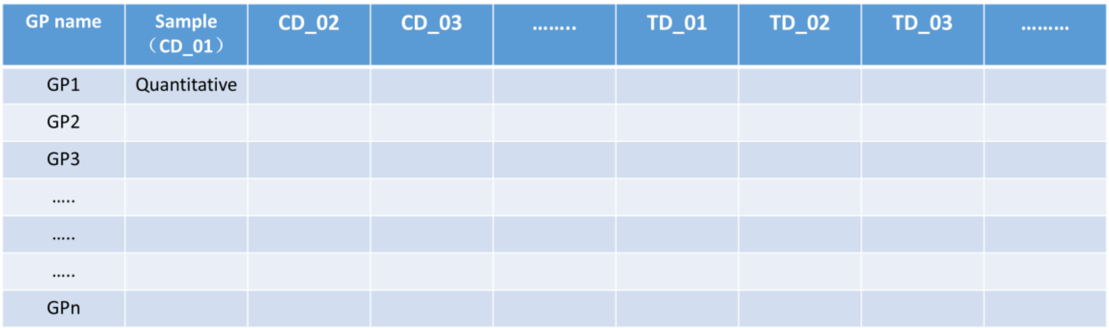
1 GP-Marker Usage

1.1 Input and Output file

Input file（discovery.csv and validation.csv）：Quantitative glycopeptide values corresponding to all samples that need to be classified. The file format is as follows; the file has the suffix ‘.csv’.



GP name	Sample （CD_01）	CD_02	CD_03	....	TD_01	TD_02	TD_03	....
GP1	Quantitative							
GP2								
GP3								
.....								
.....								
.....								
GPn								

PS：1）GP：Glycopeptide, CD: control data, TD: tumor data

2）The number starts from 01, and the single-digit number is also numbered with two digits.

3）There are no other redundant column names. Make sure the column names are in a format like CD_01.

4) Similar samples have the same name in the discovery set and validation set. For example, if they are both control samples, they are all named CD.

5) Machine learning classification only supports two-class classification problems for the time being.

1.2 Output file:

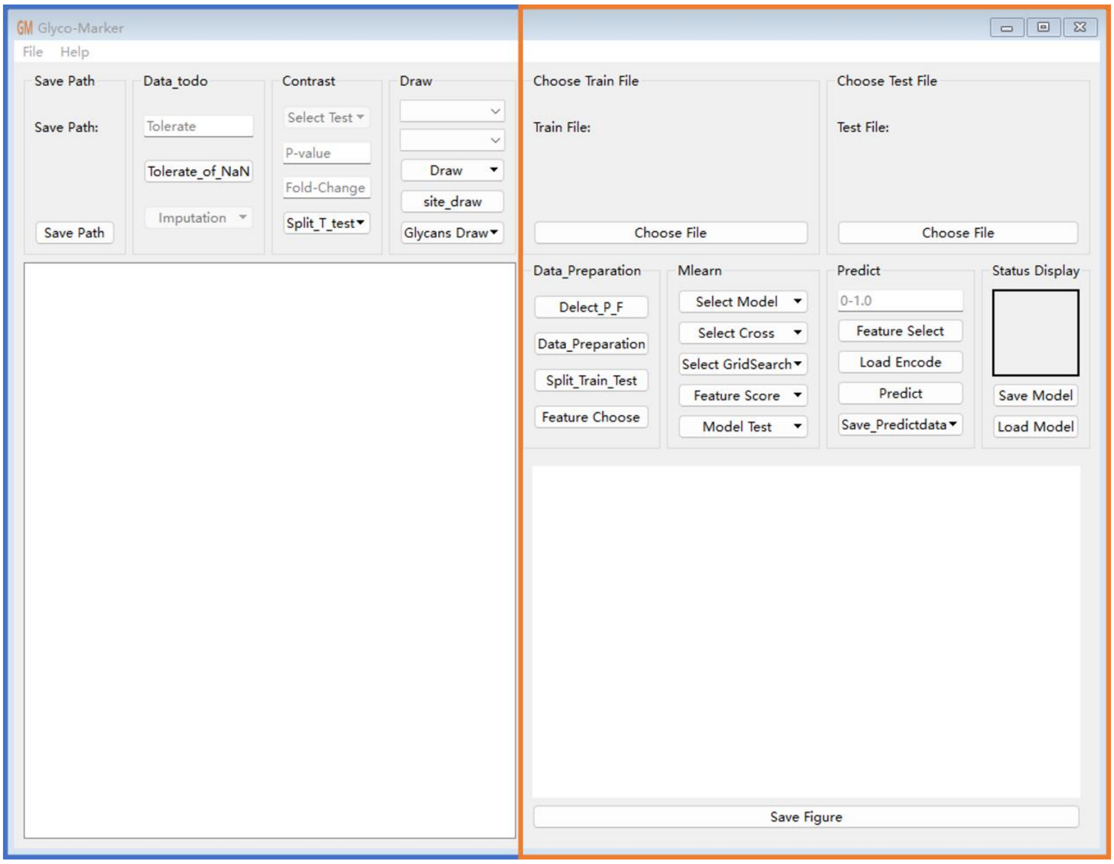
- (T-test).csv: The name of the file is selected by the user and contains the P-value and Fold-change values of the features whose missing value ratio meets the requirements.
- MLDATA.csv: The file contains features that satisfy P-value and Fold-change card values. This file is used for subsequent machine learning.
- randomfeature.csv: This file contains the contribution of each feature to the model establishment.
- all_auc.csv: This file contains the AUC values calculated for all features based on the data.
- predictdata.csv: This file contains the specific prediction results for each sample of the validation set.
- (RandomForestClassifier)_model.joblib: The file name is related to the selected machine

learning model and is a model saved by the user for next call.

- (RandomForestClassifier)_model_encoder.pkl: The file name is also related to the selected machine learning model and is the correspondence file between the sample name and the learning parameters.

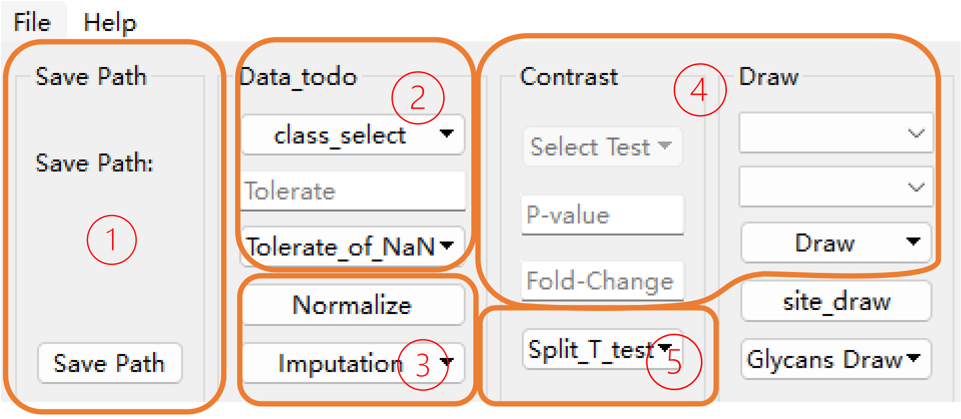
PS: The marked part of the string is a name that changes based on user input, and other file names are fixed names.

2 The process of GP-Marker



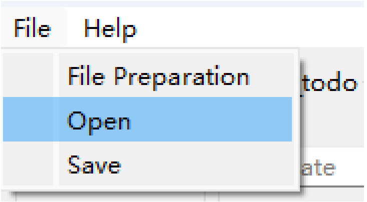
The main window of GP-Marker: 1) The blue box area corresponds to the missing value processing and T-test module. 2) The orange box area corresponds to the machine learning model training and prediction module.

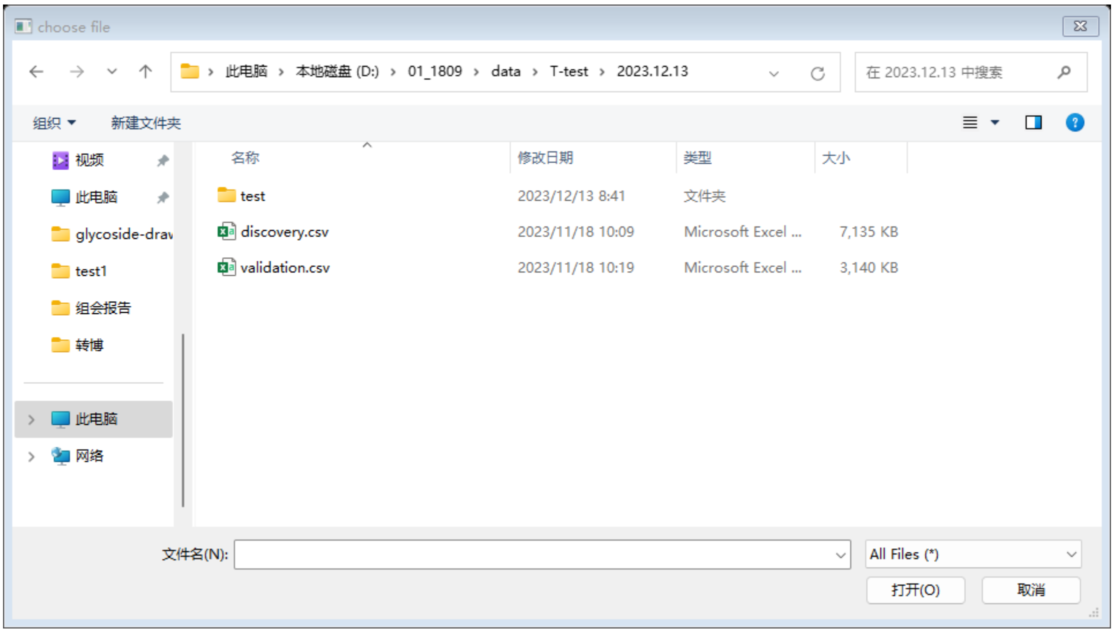
2.1 Missing value processing and T-test



If your file format is consistent with our requirements in part one, you can directly follow the steps below

1) Click ‘open’ in ‘File’ to bring up the file selection window and select ‘discovery.csv’



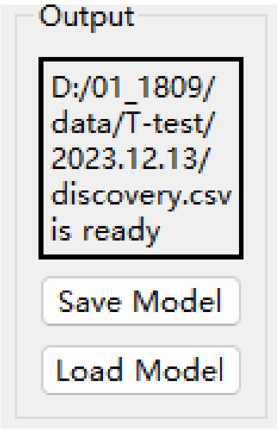

choose file

此电脑 › 本地磁盘 (D:) › 01_1809 › data › T-test › 2023.12.13

名称	修改日期	类型	大小
test	2023/12/13 8:41	文件夹	
discovery.csv	2023/11/18 10:09	Microsoft Excel ...	7,135 KB
validation.csv	2023/11/18 10:19	Microsoft Excel ...	3,140 KB

文件名(N):

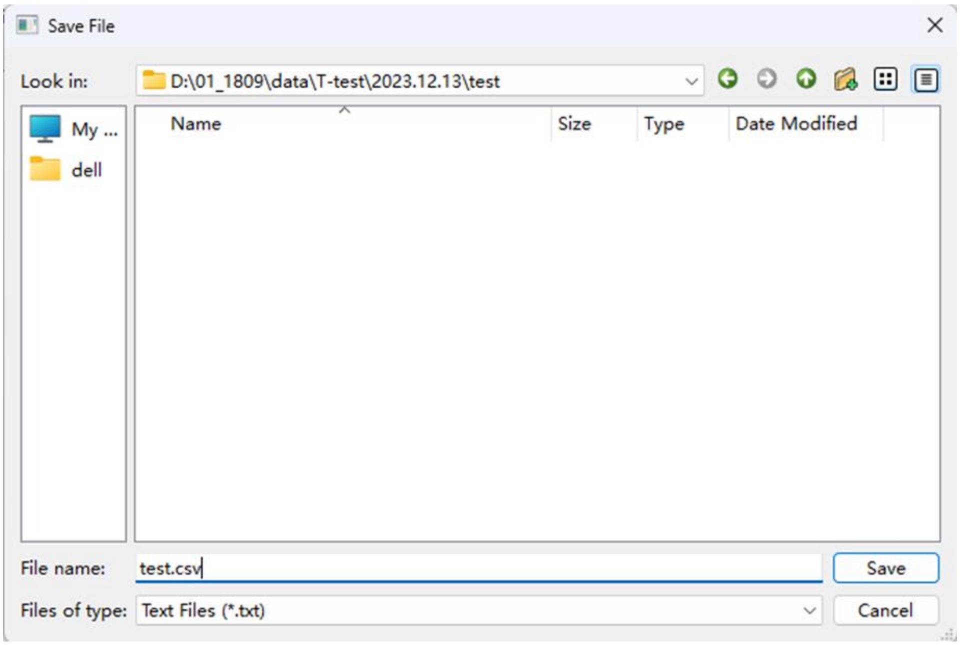
All Files (*)


Output

D:/01_1809/data/T-test/2023.12.13/discovery.csv is ready

If the reading is successful, it will be displayed in a small window.

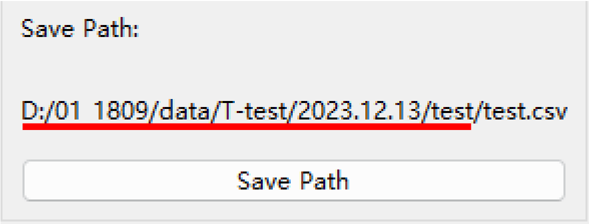
2）Click ‘Save Path’ in Part 1 to bring up the path window and enter a save path in csv format.



Look in: D:\01_1809\data\T-test\2023.12.13\test

File name: test.csv

Files of type: Text Files (*.txt)

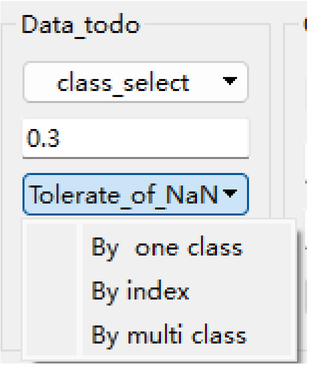


Save Path:

D:/01_1809/data/T-test/2023.12.13/test/test.csv

Show save path here.
PS: Note that subsequent files will be saved in this folder path.

3）Filter missing values in Part 2



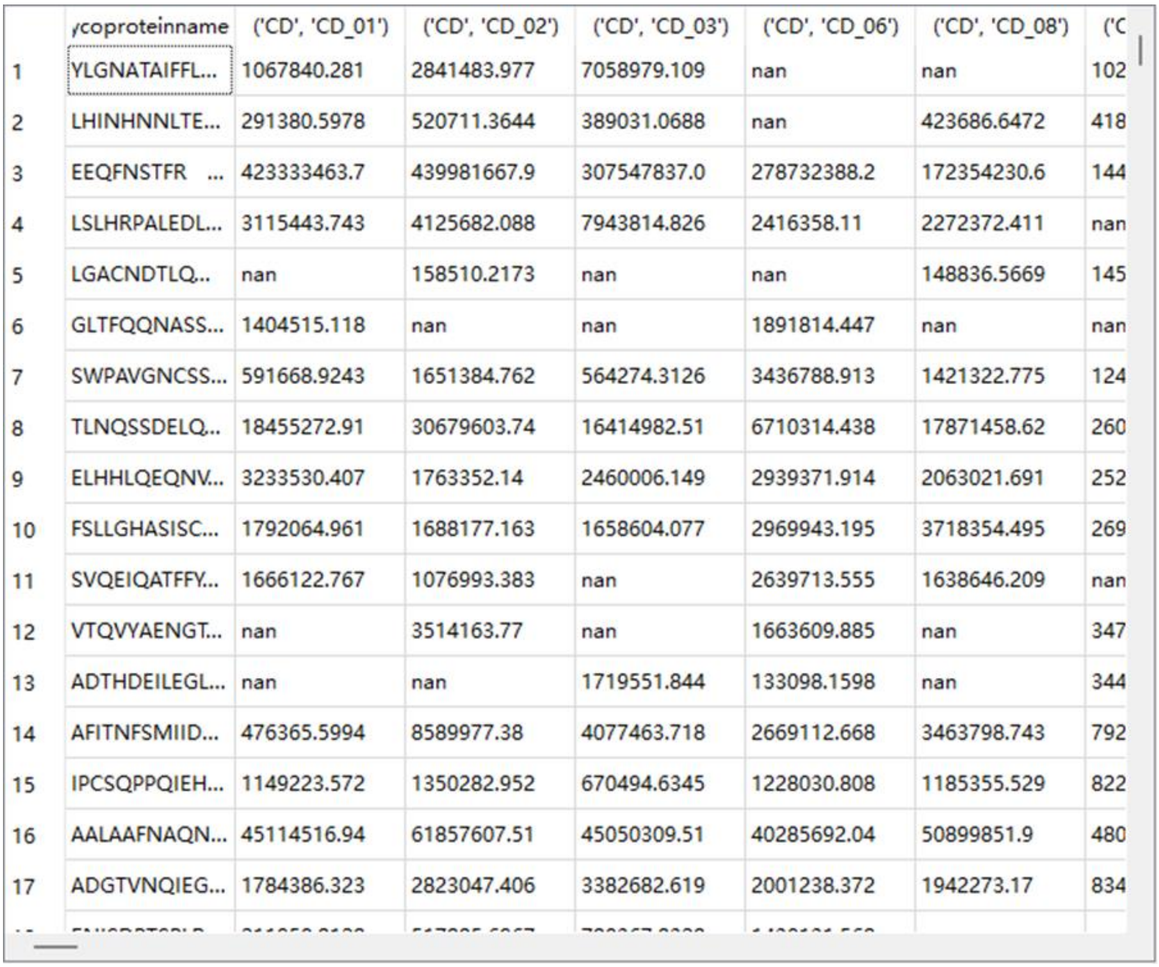
0.3

Enter the allowed range of missing values and click ‘Tolerate_of_NaN’

By one class： The proportion of missing values in one category is less than 0.3

By index： The total missing value ratio of a feature is less than 0.3

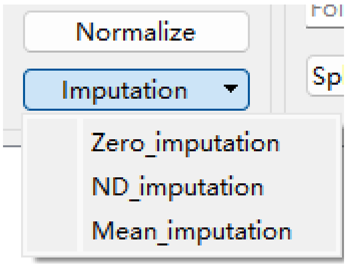
By multi class： The proportion of missing values in each category is less than 0.3



	ycoproteinname	('CD', 'CD_01')	('CD', 'CD_02')	('CD', 'CD_03')	('CD', 'CD_06')	('CD', 'CD_08')	('C
1	YLGNATAIFFL...	1067840.281	2841483.977	7058979.109	nan	nan	102
2	LHINHNNLTE...	291380.5978	520711.3644	389031.0688	nan	423686.6472	418
3	EEQFNSTFR ...	423333463.7	439981667.9	307547837.0	278732388.2	172354230.6	144
4	LSLHRPALEDL...	3115443.743	4125682.088	7943814.826	2416358.11	2272372.411	nan
5	LGACNDTLQ...	nan	158510.2173	nan	nan	148836.5669	145
6	GLTFQQNASS...	1404515.118	nan	nan	1891814.447	nan	nan
7	SWPAVGNCSS...	591668.9243	1651384.762	564274.3126	3436788.913	1421322.775	124
8	TLNQSSDELQ...	18455272.91	30679603.74	16414982.51	6710314.438	17871458.62	260
9	ELHHLQEQNV...	3233530.407	1763352.14	2460006.149	2939371.914	2063021.691	252
10	FSLLGHASISC...	1792064.961	1688177.163	1658604.077	2969943.195	3718354.495	269
11	SVQEIQATFFY...	1666122.767	1076993.383	nan	2639713.555	1638646.209	nan
12	VTQVYAENGT...	nan	3514163.77	nan	1663609.885	nan	347
13	ADTHDEILEGL...	nan	nan	1719551.844	133098.1598	nan	344
14	AFITNFSMIID...	476365.5994	8589977.38	4077463.718	2669112.668	3463798.743	792
15	IPCSQPPQIEH...	1149223.572	1350282.952	670494.6345	1228030.808	1185355.529	822
16	AALAAFNAQN...	45114516.94	61857607.51	45050309.51	40285692.04	50899851.9	480
17	ADGTVNQIEG...	1784386.323	2823047.406	3382682.619	2001238.372	1942273.17	834

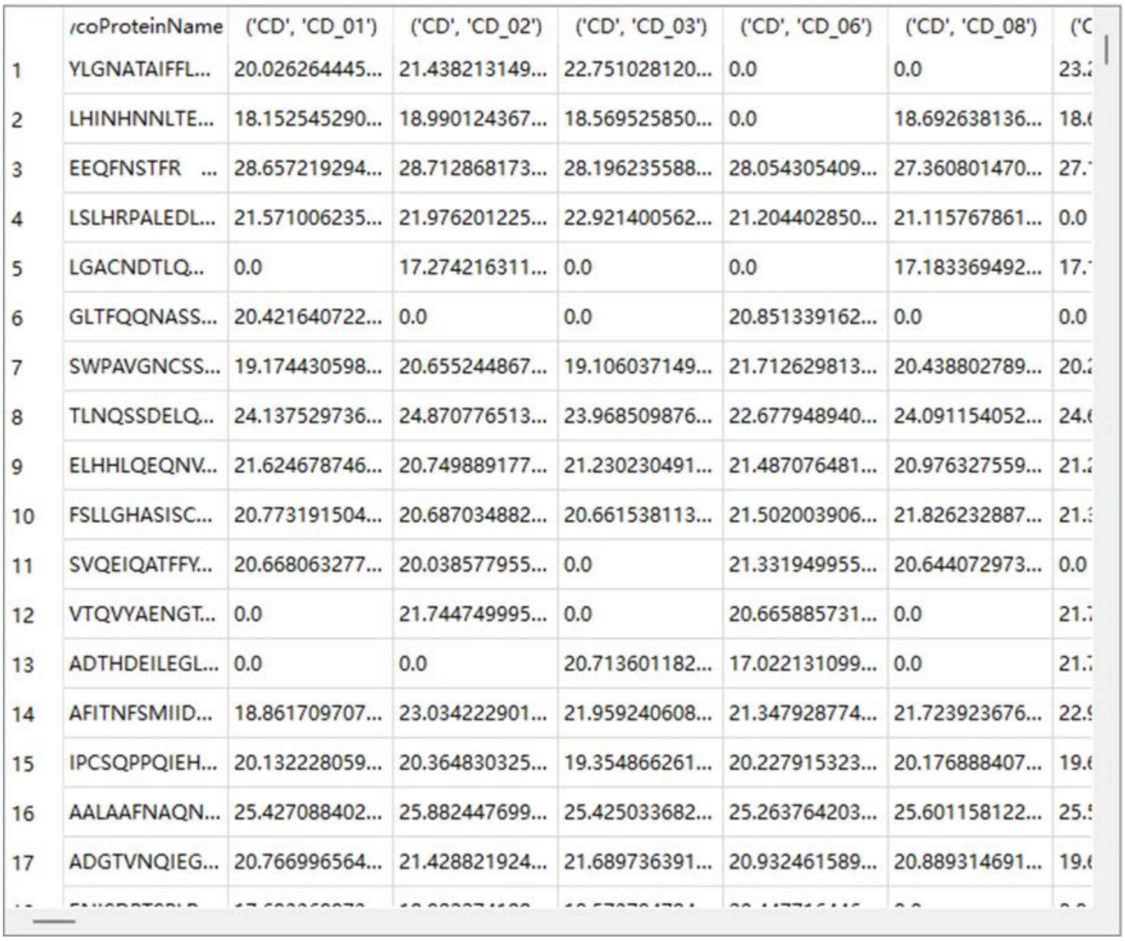
The display box displays the data after filtering missing values.

4) Impute missing values in Part 3



Normalize:
Use the median value of the first sample as the standard to normalize other sample data.

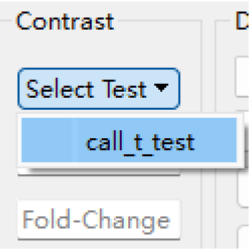
Imputation:
1 Fill missing values with 0 (take supplementing 0 as an example)
2 Missing values are randomly filled according to the left-skewed kurtosis normal distribution.
3 Fill missing values with mean



	/coProteinName	('CD', 'CD_01')	('CD', 'CD_02')	('CD', 'CD_03')	('CD', 'CD_06')	('CD', 'CD_08')	('C
1	YLGNATAIFFL...	20.026264445...	21.438213149...	22.751028120...	0.0	0.0	23.
2	LHINHNNLTE...	18.152545290...	18.990124367...	18.569525850...	0.0	18.692638136...	18.
3	EEQFNSTFR ...	28.657219294...	28.712868173...	28.196235588...	28.054305409...	27.360801470...	27.
4	LSLHRPALEDL...	21.571006235...	21.976201225...	22.921400562...	21.204402850...	21.115767861...	0.0
5	LGACNDTLQ...	0.0	17.274216311...	0.0	0.0	17.183369492...	17.
6	GLTFQQNASS...	20.421640722...	0.0	0.0	20.851339162...	0.0	0.0
7	SWPAVGNCSS...	19.174430598...	20.655244867...	19.106037149...	21.712629813...	20.438802789...	20.
8	TLNQSSDELQ...	24.137529736...	24.870776513...	23.968509876...	22.677948940...	24.091154052...	24.
9	ELHHLQEQNV...	21.624678746...	20.749889177...	21.230230491...	21.487076481...	20.976327559...	21.
10	FSLLGHASISC...	20.773191504...	20.687034882...	20.661538113...	21.502003906...	21.826232887...	21.
11	SVQEIQATFFY...	20.668063277...	20.038577955...	0.0	21.331949955...	20.644072973...	0.0
12	VTQVYAENGT...	0.0	21.744749995...	0.0	20.665885731...	0.0	21.
13	ADTHDEILEGL...	0.0	0.0	20.713601182...	17.022131099...	0.0	21.
14	AFITNFSMIID...	18.861709707...	23.034222901...	21.959240608...	21.347928774...	21.723923676...	22.
15	IPCSQPPQIEH...	20.132228059...	20.364830325...	19.354866261...	20.227915323...	20.176888407...	19.
16	AALAAFNAQN...	25.427088402...	25.882447699...	25.425033682...	25.263764203...	25.601158122...	25.
17	ADGTVNQIEG...	20.766996564...	21.428821924...	21.689736391...	20.932461589...	20.889314691...	19.

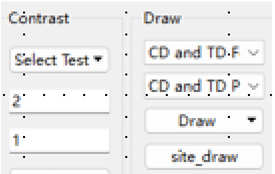
The display box is updated to the data after supplementing 0.

5) Take the T-test in Part 4

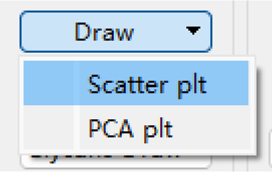


Click 'Select Test'
'call_t_test': Calculate P-value and Fold-Change

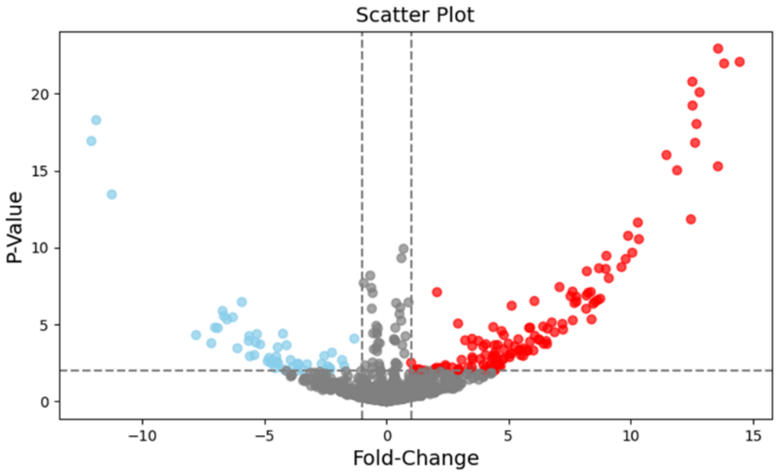
This operation will generate the first result file (T-test).csv



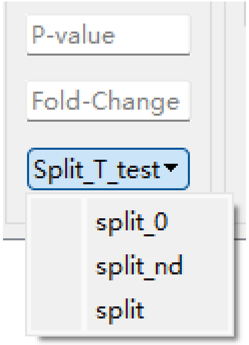
Enter the card value of P-value and Fold-change
P-value taken -log10
Fold-Change taken log2



Click 'Scatter plt' under 'Draw' to draw a volcano map.



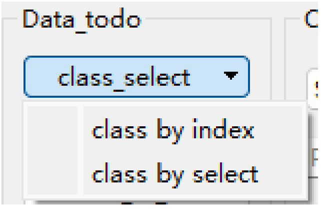
6) Generate machine learning read files in Part 5

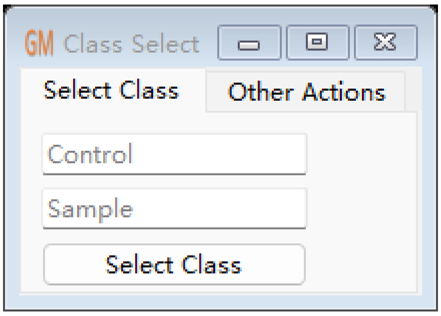


Under 'Split_T_test', select accordingly according to the value complement method.
Choose 'split'

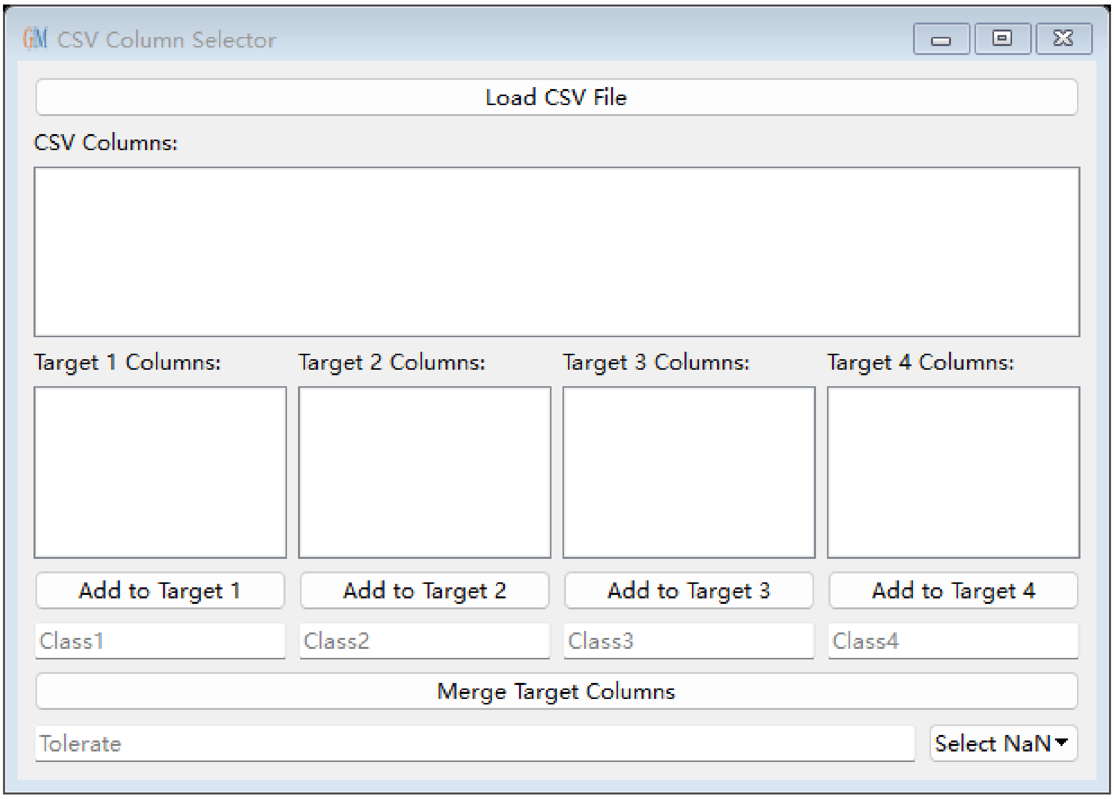
This operation will generate a second file MLDATA.csv

If your file format is inconsistent with the first part, automatic classification cannot be completed. Here are two classification methods:





'class by index': Your column name includes the characters used for classification, then fill in the two types of characteristic characters in the small window on the right.



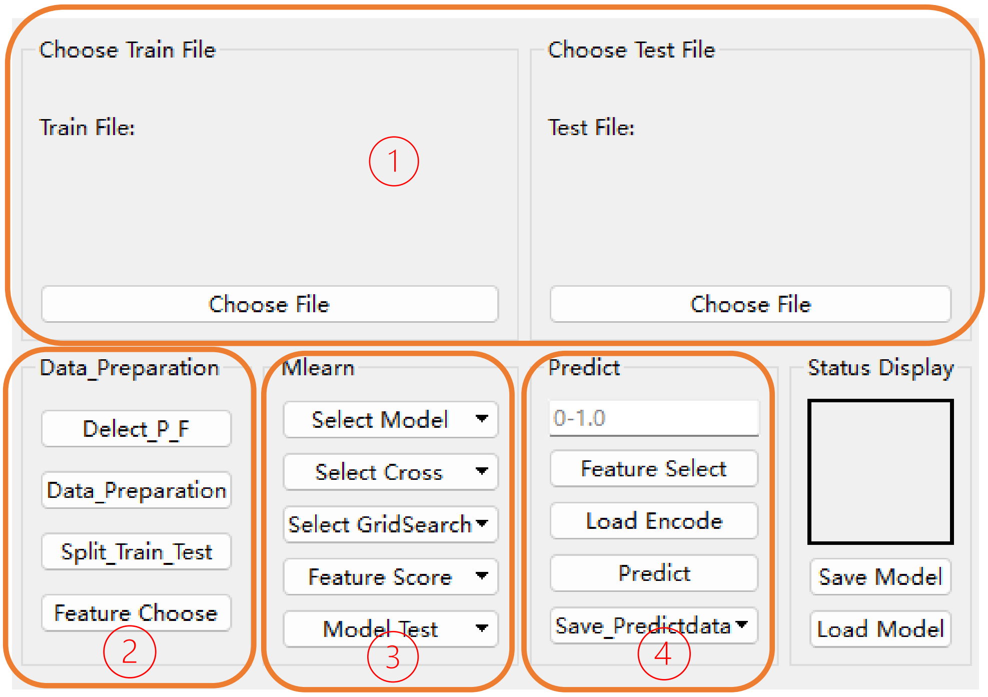
'class by select':

After selecting the file, the column names will be displayed in 'CSV columns'

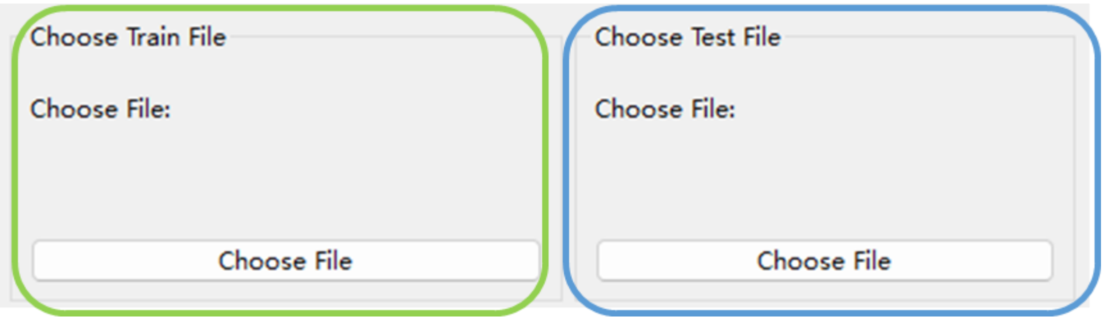
Then you can choose the column names for classification. This window supports up to four categories, and name the categories respectively, and then click 'Merge Target Columns'

And complete the filtering of missing values in this window

2.2 Machine learning model training and Prediction

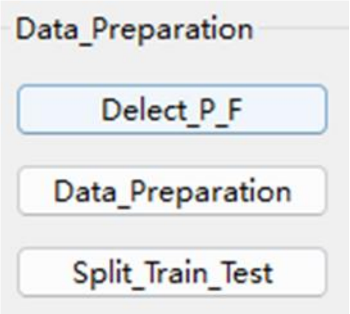


1)Machine learning file preparation in Part 1



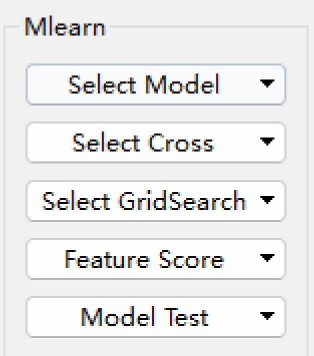
Click 'Choose File' and select MLDATA.csv and 'verification set file: 'validation.csv', respectively.

2)Data preprocessing in Part 2



Click the three buttons in sequence
Complete data sorting
And split the training set data into train and test

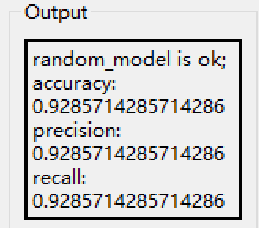
3) Machine learning model training in Part 3



'Select Model': Choose machine learning model
1 log_model
2 random_model
3 tree_model
4 svc_model
5 nb_model
6 kn_model

'Select Corss': Select the corresponding model for cross-validation.
'Select Cross': Select the corresponding model to perform grid search to optimize parameters.
'Feature Score': Give the contribution of each feature (nb_model and kn_model do not have this feature)
After using 'Feature Score' function, file 'randomfeature.csv' will be generated.
'Model Test': Provides multiple methods to measure the effectiveness of machine learning models.

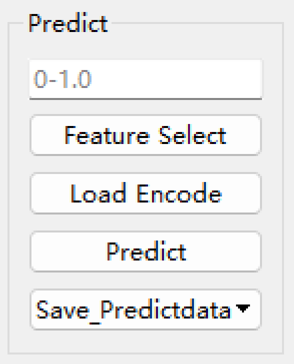
If you just want to train the model according to simple parameters, just use 'Select Model' and 'Feature Score' functions.

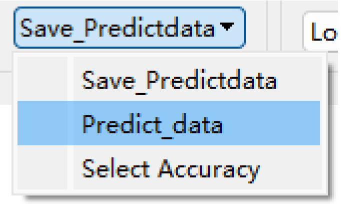


```
random_model is ok;
accuracy:
0.9285714285714286
precision:
0.9285714285714286
recall:
0.9285714285714286
```

After completing model training, information such as accuracy will be displayed here.

4) Validation set predictions in Part 4





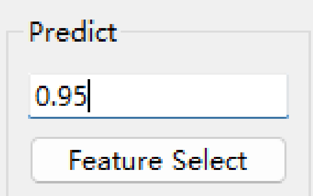
Then click Predict_data’ to generate the prediction result file: predictdata.csv

Running to this step completes the training of a simple machine learning model.

In many cases, not performing feature screening will result in hundreds of features being used for model training, while disease marker screening often only focuses on a few more obvious features, so the software also provides a series of feature screening methods.

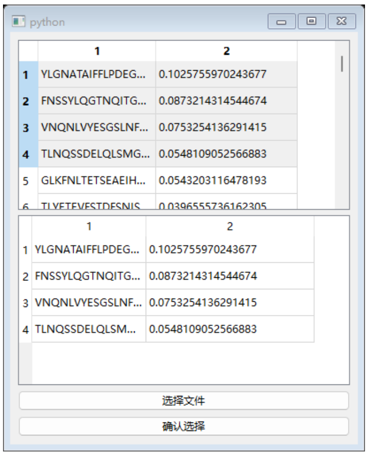
2.3 Feature filtering operation

1) Optimize features in batches based on the sum of contributions (In Part 5 of section 2.2)



Enter the sum of retained contributions
Click ‘Feature Select’
Automatically filter these features in the ‘(TD_CD).csv’ file and overwrite the file
PS: The sum of feature contributions is 1.0

2) Autonomous selection of feature training models (Click ‘Feature Choose’ in Part 2 of section 2.2)



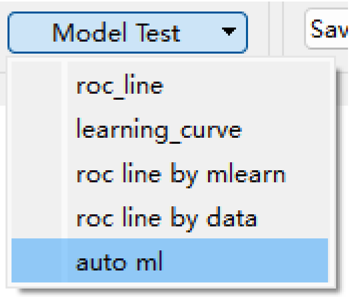
	1	2
1	YLGNATAIFFLPDEG...	0.1025755970243677
2	FNSSYLQGTNQITG...	0.0873214314544674
3	VNQNLVYESGSLNF...	0.0753254136291415
4	TLNQSSDELQLSMG...	0.0548109052566883
5	GLKFNLTETSEAEIH...	0.0543203116478193
6	TLYETEVESTDESNIS	0.0396555736162305

	1	2
1	YLGNATAIFFLPDEG...	0.1025755970243677
2	FNSSYLQGTNQITG...	0.0873214314544674
3	VNQNLVYESGSLNF...	0.0753254136291415
4	TLNQSSDELQLSM...	0.0548109052566883

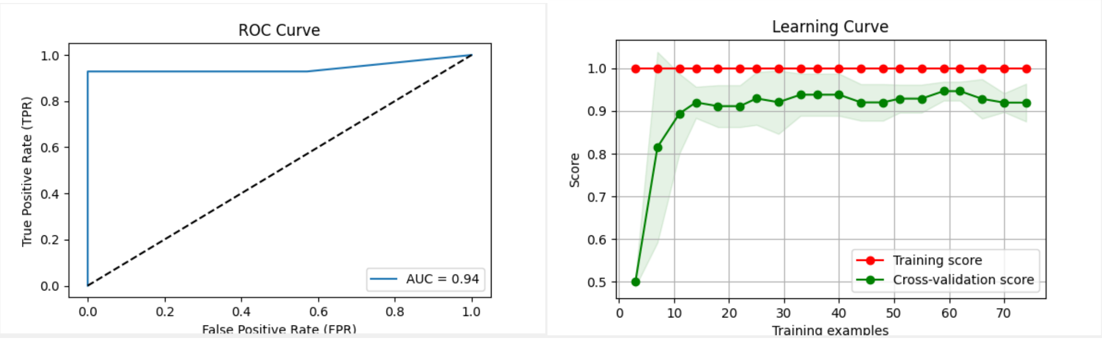
To select the file here, you need to select the ‘randomfeature.csv’ file.
Select the features of interest in the upper box and click to confirm
Overwrite ‘MLDATA.csv’ file

2.4 Other function of GP-Marker

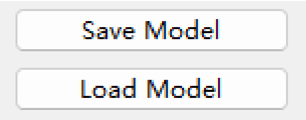
1) Related model performance curves



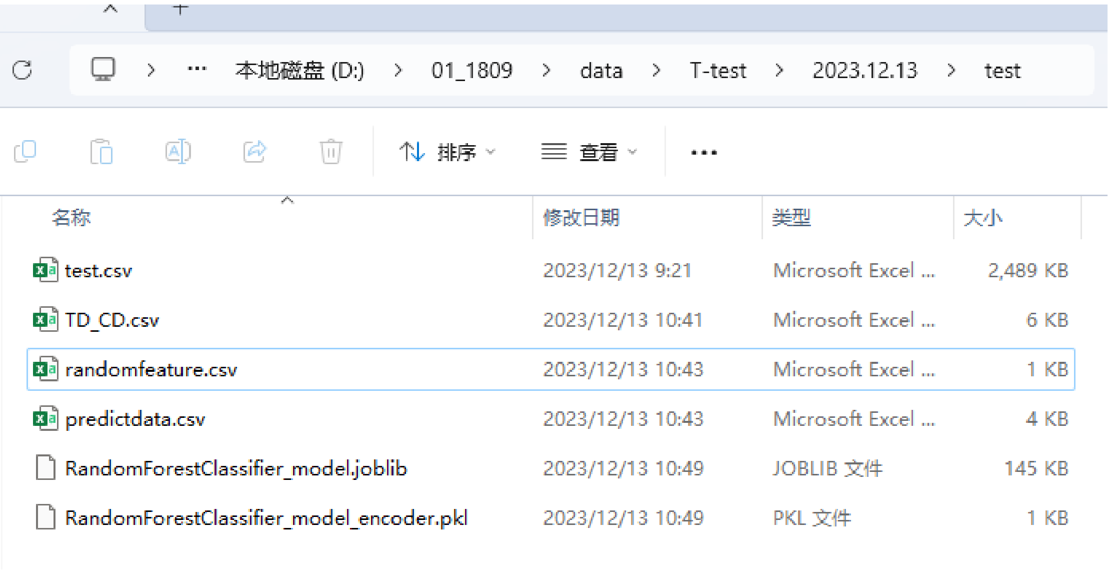
Provides drawing model curves:
'roc_line': draw ROC curve
'learning_curve': draw learning curve



2) Save Model



save model button
load model button



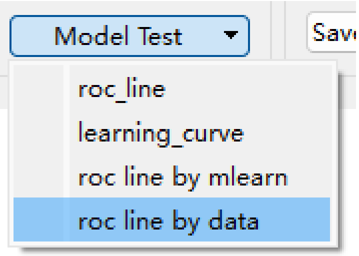
名称	修改日期	类型	大小
test.csv	2023/12/13 9:21	Microsoft Excel ...	2,489 KB
TD_CD.csv	2023/12/13 10:41	Microsoft Excel ...	6 KB
randomfeature.csv	2023/12/13 10:43	Microsoft Excel ...	1 KB
predictdata.csv	2023/12/13 10:43	Microsoft Excel ...	4 KB
RandomForestClassifier_model.joblib	2023/12/13 10:49	JOBLIB 文件	145 KB
RandomForestClassifier_model_encoder.pkl	2023/12/13 10:49	PKL 文件	1 KB

Files ending in '.joblib' are model files.

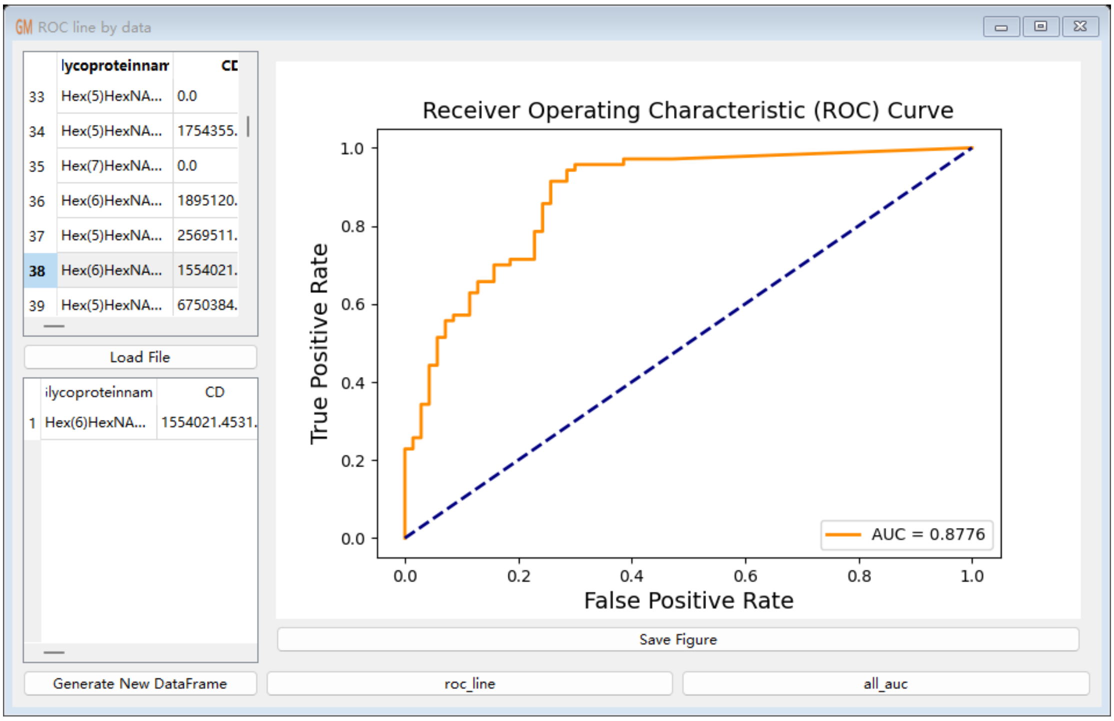
Files ending in '.pkl' are prediction mode files.

This model can then be used directly to call.

3) Draw the ROC curve of a single glycopeptide



Click the ‘roc line by data’
New window will be opened



	lycoproteinnam	CD
33	Hex(5)HexNA...	0.0
34	Hex(5)HexNA...	1754355.
35	Hex(7)HexNA...	0.0
36	Hex(6)HexNA...	1895120.
37	Hex(5)HexNA...	2569511.
38	Hex(6)HexNA...	1554021.
39	Hex(5)HexNA...	6750384.

	ilycoproteinnam	CD
1	Hex(6)HexNA...	1554021.4531.

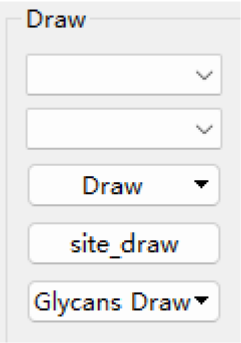
Load File: MLDATA.csv

Select one or more glycopeptides of interest in the upper box, and then click ‘roc_line’ to draw the curve.

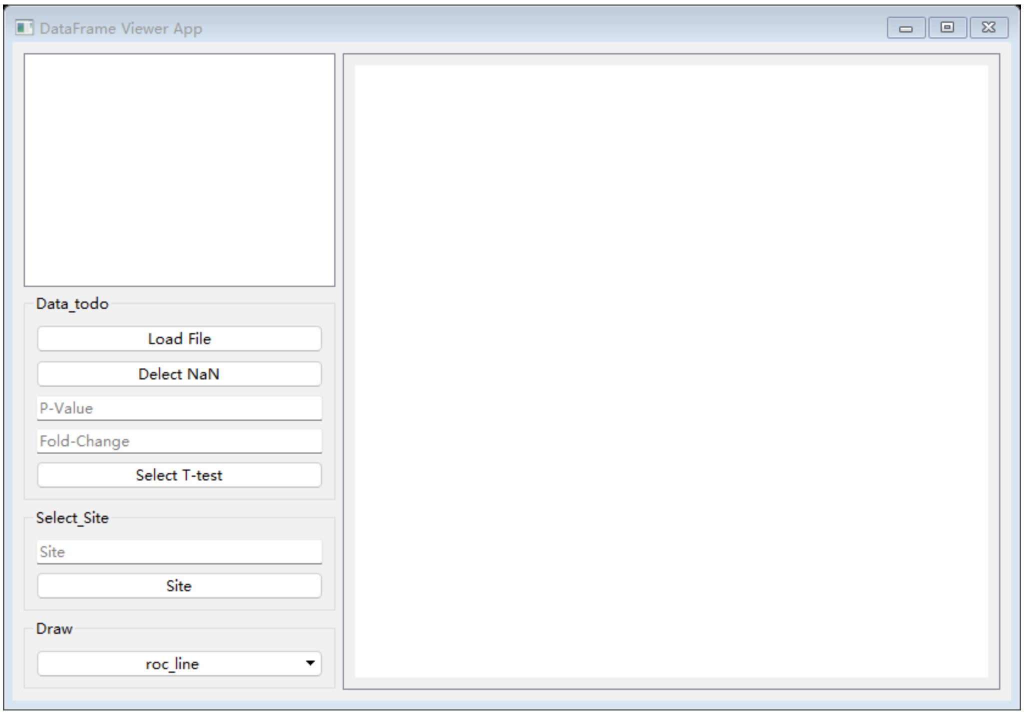
You can also click ‘all_auc’, an AUC value file corresponding to each feature will be generated.

PS: all_auc.csv

4) Display site glycoform distribution function



Click the ‘site_draw’
New window will be opened



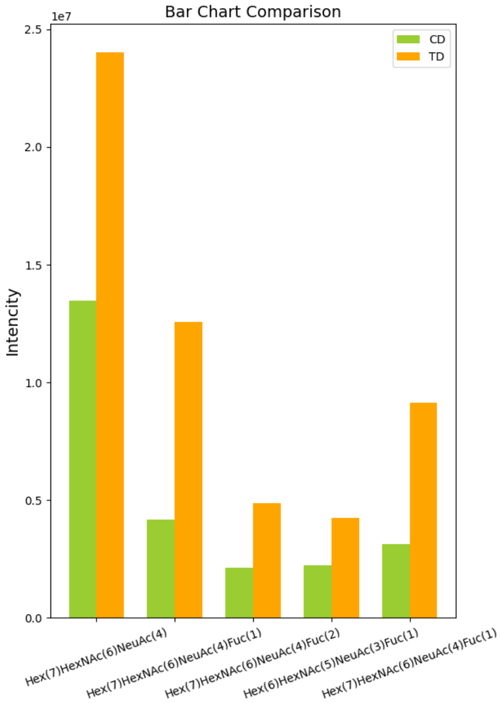
Load File: discovery.csv

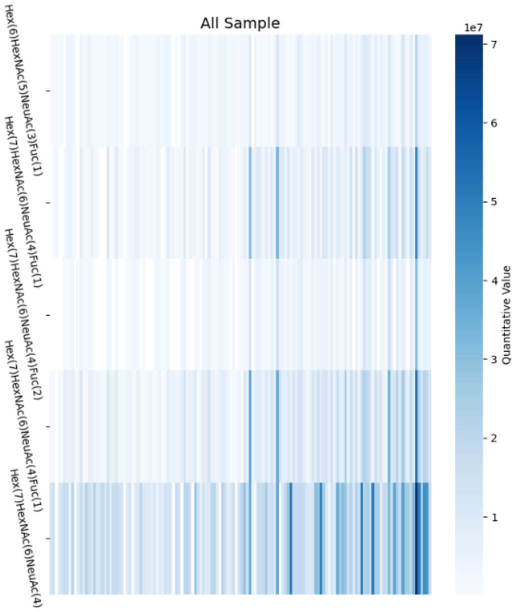
The canvas on the right shows the drawn distribution map

Provides methods for filtering missing values

Provides T-test card P-value and Fold-Change methods

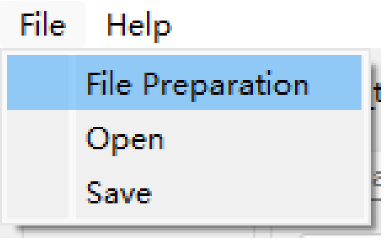
Two comparison charts are provided: bar chart and heat map



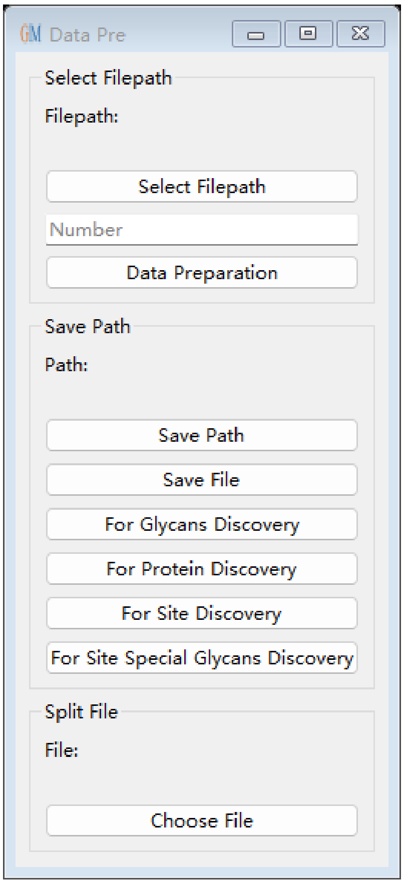


5) Auxiliary format handling

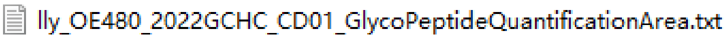
GP-Marker can directly count the quantitative result file output by Glyco-Decipher as discovery.csv.



Click the ‘File Preparation’
New window will be opened



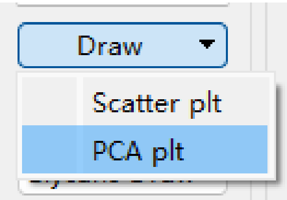
‘Select Filepath’: Select the quantitative results folder of Glyco-Decipher.



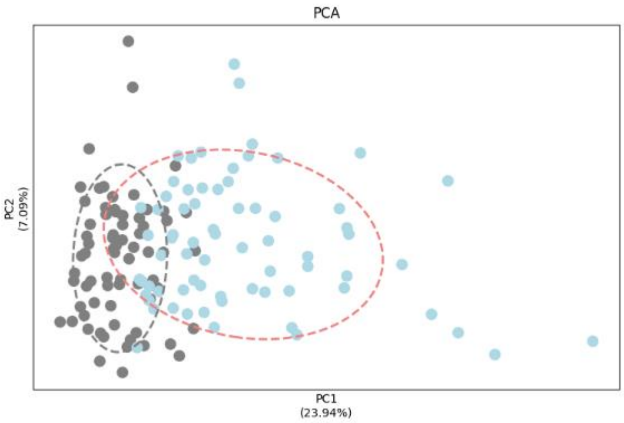
lly_OE480_2022GCHC_CD01_GlycoPeptideQuantificationArea.txt

‘Data Preparation’: Automatically count all files ending with ‘Area’
And generate csv file format files recognized by the software
PS: Before using this function, enter the number of the underscore partition in the sample name in the ‘Number’ box. For example, enter 4 for the above file. If you do not enter a number, the file name will be used as the column name of this sample.
‘Save Path’: Select save path
‘Save File’: output discovery.csv at the intact glycopeptide level
‘For Glycans Discovery’: output discovery.csv at the glycans level
‘For Protein Discovery’: output discovery.csv at the protein level
‘For Site Discovery’: output discovery.csv at the site level
‘For Site Special Glycans Discovery’: output discovery.csv at the site special glycans level
PS: Before counting data at other levels, you need to count the statistics of complete glycopeptide data. If you have the discovery.csv at the intact glycopeptide level, you can use the ‘Choose File’ button to import the file and then choose the different levels of splitting methods above.

6) Draw PCA

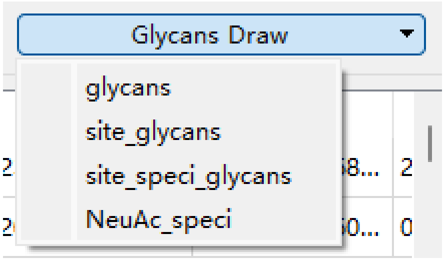


Click the ‘PCA plt’
Draw PCA

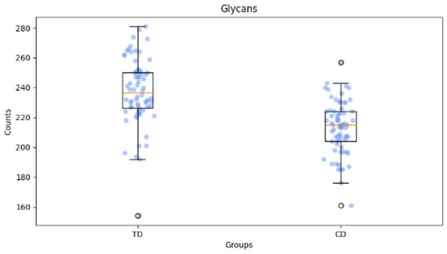


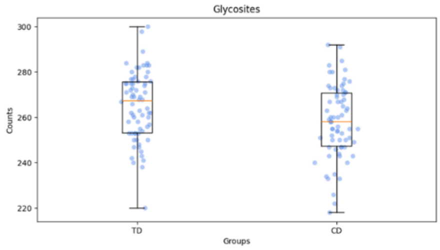
Before drawing PCA, you need to complete the screening of missing values and card P-value values.

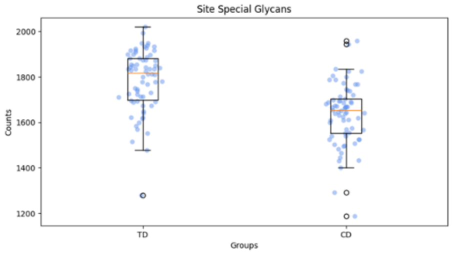
7) Plot multi-level distributions

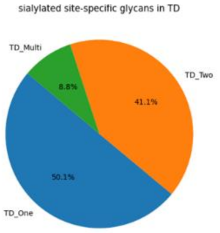


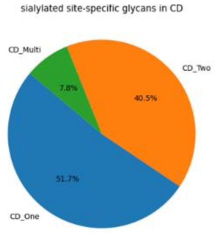
Four types of drawings under Glycans Draw
Glycan level quantity distribution
Site level quantity distribution
Site-specific glycans number distribution
Sialic acid distribution level





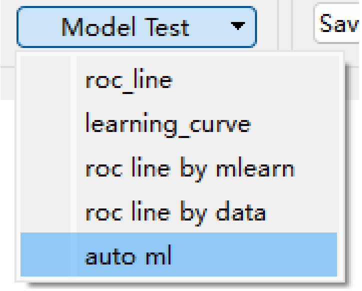




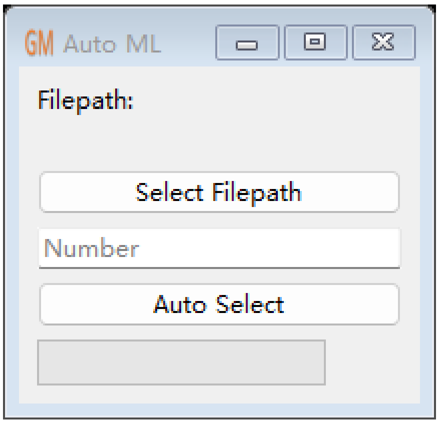


This function can be used after importing discovery.csv

8) Screen the best combination of features



Choose ‘auto ml’
Open a new window.



‘Select Filepath’: MLDATA.csv
‘Number’: Select the number of combined features
‘Auto Select’: Iterate over all possible combinations to train the random forest model and output the accuracy result file.