La base de datos empleada para la realización del modelo se obtuvo del artículo “Using a Graph Convolutional Neural Network Model to Identify Bile Salt Export Pump Inhibitors”, y se puede acceder <https://doi.org/10.1021/acsomega.3c01583>. En este artículo se recopila una base de datos de 925 compuestos a partir de la herramienta web BindingDB compuesta por 152 BSEP inhibitors y 772 non-inhibitors.

We collected 1,689 compounds with publicly available BSEP bioactivity data from BindingDB (accessed on 5-19-2022).31 BindingDB curates the activity data for each target from various literature sources, including patents.31 We used Pipeline Pilot (Version18.1.100.11) to preprocess the molecules, removing duplicate compounds, salts, and mixtures and standardizing themolecules.32 Standardization refers to a molecule preprocessing step wherein proper bond order, aromaticity, and hydrogens are assigned.33 As suggested in the International Transporter Consortium workflow on BSEP inhibition in drug discovery, we used a half-maximal inhibitory concentration(IC50) cutoff of 25 μM,4 designating compounds with IC50values <25 μM as inhibitors and >100 μM as noninhibitors of BSEP. We excluded compounds with IC50 values between 25and 100 μM from our analysis. Our final dataset consisted of925 compounds with 152 BSEP inhibitors and 773 non-inhibitors

No se requirió tratado inicial de los datos, ya que la base de datos constaba de dos columnas (SMILES y BSEP ACTIVITY) con la información. BSEP ACTIVITY se cambio como “y”, además contenía una variable binomial para clasificar si un compuesto era inhibidor o no (1 = inhibidor, 0 = no inhibidor). Tras procesar los datos con HYGEIA no se eliminó ningún compuesto.

Para el cálculo de descriptores en Wotan se utilizaron los siguientes modelos:

    'AtomCentred',

    'Autocorrelation',

    'BurdenEigenvalues',

    'ConnectivityIndices',

    'Constitutional',

    'EdgeAdjacency',

    'Eigenvalues',

    # 'GETAWAY', #3D

    'InformationIndices',

    # 'Rdkit3D', #3D

    'Topological',

    'TopologicalCharge',

    'TwoDimensional',

    'WalkPathCounts',

    'FunctionalGroup',

    'MoeType', #some3D

    # 'CPSA', #some3D

    'EState', #some3D

    'SLogP', #some3D

A continuación se muestran las opciones seleccionadas en NEO:

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 1

This part of the code will do the y" transformation, randomization of the dataset order and Knn imputation.

From this version of NEO is its compulsory to perform the inputation here, as this will create the un-imputed file needed for reimputation.

[+] "y" transformation

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/" folder is needed

This file must be called: "BSEP\_NO3D\_local-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?

Please select your type of model:

[1] Regression

[2] Classification

Your choice: 2

I am so sorry, there is nothing yet for your request. Please try tomorrow with more coffee and cookies.

[+] dataset random sort

The following file has been created (save it as you will need it for feature reduction):

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-calculated\_preimputation.csv

[+] dataset imputation

Size of the database, preimputation: (925, 3722)

Warning! Some of descriptors have too many NaN values and have been removed: ['LPRS', 'VDA', 'MDDD', 'MAXsLi', 'MAXssBe', 'MAXssssBe', 'MAXssBH', 'MAXsssB', 'MAXssssB', 'MAXsCH3', 'MAXdCH2', 'MAXssCH2', 'MAXtCH', 'MAXdsCH', 'MAXaaCH', 'MAXsssCH', 'MAXddC', 'MAXtsC', 'MAXdssC', 'MAXaasC', 'MAXaaaC', 'MAXssssC', 'MAXsNH3', 'MAXsNH2', 'MAXssNH2', 'MAXdNH', 'MAXssNH', 'MAXaaNH', 'MAXtN', 'MAXsssNH', 'MAXdsN', 'MAXaaN', 'MAXsssN', 'MAXddsN', 'MAXaasN', 'MAXssssN', 'MAXsOH', 'MAXdO', 'MAXssO', 'MAXaaO', 'MAXsF', 'MAXsSiH3', 'MAXssSiH2', 'MAXsssSiH', 'MAXssssSi', 'MAXsPH2', 'MAXssPH', 'MAXsssP', 'MAXdsssP', 'MAXsssssP', 'MAXsSH', 'MAXdS', 'MAXssS', 'MAXaaS', 'MAXdssS', 'MAXddssS', 'MAXsCl', 'MAXsGeH3', 'MAXssGeH2', 'MAXsssGeH', 'MAXssssGe', 'MAXsAsH2', 'MAXssAsH', 'MAXsssAs', 'MAXsssdAs', 'MAXsssssAs', 'MAXsSeH', 'MAXdSe', 'MAXssSe', 'MAXaaSe', 'MAXdssSe', 'MAXddssSe', 'MAXsBr', 'MAXsSnH3', 'MAXssSnH2', 'MAXsssSnH', 'MAXssssSn', 'MAXsI', 'MAXsPbH3', 'MAXssPbH2', 'MAXsssPbH', 'MAXssssPb', 'MINsLi', 'MINssBe', 'MINssssBe', 'MINssBH', 'MINsssB', 'MINssssB', 'MINsCH3', 'MINdCH2', 'MINssCH2', 'MINtCH', 'MINdsCH', 'MINaaCH', 'MINsssCH', 'MINddC', 'MINtsC', 'MINdssC', 'MINaasC', 'MINaaaC', 'MINssssC', 'MINsNH3', 'MINsNH2', 'MINssNH2', 'MINdNH', 'MINssNH', 'MINaaNH', 'MINtN', 'MINsssNH', 'MINdsN', 'MINaaN', 'MINsssN', 'MINddsN', 'MINaasN', 'MINssssN', 'MINsOH', 'MINdO', 'MINssO', 'MINaaO', 'MINsF', 'MINsSiH3', 'MINssSiH2', 'MINsssSiH', 'MINssssSi', 'MINsPH2', 'MINssPH', 'MINsssP', 'MINdsssP', 'MINsssssP', 'MINsSH', 'MINdS', 'MINssS', 'MINaaS', 'MINdssS', 'MINddssS', 'MINsCl', 'MINsGeH3', 'MINssGeH2', 'MINsssGeH', 'MINssssGe', 'MINsAsH2', 'MINssAsH', 'MINsssAs', 'MINsssdAs', 'MINsssssAs', 'MINsSeH', 'MINdSe', 'MINssSe', 'MINaaSe', 'MINdssSe', 'MINddssSe', 'MINsBr', 'MINsSnH3', 'MINssSnH2', 'MINsssSnH', 'MINssssSn', 'MINsI', 'MINsPbH3', 'MINssPbH2', 'MINsssPbH', 'MINssssPb'] The limit is marked to the 15.0 %

[+] fitting

[+] transforming

Size of the database, postimputation: (925, 3561)

C-001 C-002 C-003 C-004 C-005 C-006 C-007 ... SsI SsPbH3 SssPbH2 SsssPbH SssssPb SLogP SMR

0 1.0 0.0 0.0 0.0 0.0 4.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 4.19490 88.9710

1 2.0 0.0 0.0 0.0 0.0 2.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.56360 51.4885

2 1.0 1.0 0.0 0.0 1.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.85510 67.2384

3 1.0 1.0 0.0 1.0 3.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.77390 78.3157

4 1.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 2.54542 71.0375

.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

920 2.0 7.0 3.0 2.0 0.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 4.11138 93.1328

921 0.0 1.0 0.0 0.0 1.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 3.37120 83.0670

922 0.0 0.0 0.0 0.0 2.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 0.06890 45.3086

923 0.0 5.0 0.0 0.0 0.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 2.78510 54.5578

924 0.0 0.0 0.0 0.0 0.0 1.0 0.0 ... 1.68662 0.0 0.0 0.0 0.0 -1.27010 65.9383

[925 rows x 3561 columns]

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-calculated\_imputed\_ytransformed.csv

Do you want to perform any other step?(y/n):

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 2

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/" folder is needed

This file must be called: "BSEP\_NO3D\_local-calculated\_imputed\_ytransformed.csv"

Continue (Y/n)?

[1] Initial feature reduction: infinite, correlated, constant and empty values

0 infinite values

0 features with greater than 0.00 missing values.

1466 features with a correlation magnitude greater than 0.90.

1160 features with a single unique value.

Data has not been one-hot encoded

Removed 2626 features including one-hot features.

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-initial\_reduction.csv

Do you want to perform any other step?(y/n): n

Thanks for using NEO!

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 1

This part of the code will do the y" transformation, randomization of the dataset order and Knn imputation.

From this version of NEO is its compulsory to perform the inputation here, as this will create the un-imputed file needed for reimputation.

[+] "y" transformation

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/" folder is needed

This file must be called: "BSEP\_NO3D\_local-paralel\_calculated\_with\_y.csv"

Continue (Y/n)?

Please select your type of model:

[1] Regression

[2] Classification

Your choice: 2

I am so sorry, there is nothing yet for your request. Please try tomorrow with more coffee and cookies.

[+] dataset random sort

The following file has been created (save it as you will need it for feature reduction):

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-calculated\_preimputation.csv

[+] dataset imputation

Size of the database, preimputation: (925, 3722)

Warning! Some of descriptors have too many NaN values and have been removed: ['LPRS', 'VDA', 'MDDD', 'MAXsLi', 'MAXssBe', 'MAXssssBe', 'MAXssBH', 'MAXsssB', 'MAXssssB', 'MAXsCH3', 'MAXdCH2', 'MAXssCH2', 'MAXtCH', 'MAXdsCH', 'MAXaaCH', 'MAXsssCH', 'MAXddC', 'MAXtsC', 'MAXdssC', 'MAXaasC', 'MAXaaaC', 'MAXssssC', 'MAXsNH3', 'MAXsNH2', 'MAXssNH2', 'MAXdNH', 'MAXssNH', 'MAXaaNH', 'MAXtN', 'MAXsssNH', 'MAXdsN', 'MAXaaN', 'MAXsssN', 'MAXddsN', 'MAXaasN', 'MAXssssN', 'MAXsOH', 'MAXdO', 'MAXssO', 'MAXaaO', 'MAXsF', 'MAXsSiH3', 'MAXssSiH2', 'MAXsssSiH', 'MAXssssSi', 'MAXsPH2', 'MAXssPH', 'MAXsssP', 'MAXdsssP', 'MAXsssssP', 'MAXsSH', 'MAXdS', 'MAXssS', 'MAXaaS', 'MAXdssS', 'MAXddssS', 'MAXsCl', 'MAXsGeH3', 'MAXssGeH2', 'MAXsssGeH', 'MAXssssGe', 'MAXsAsH2', 'MAXssAsH', 'MAXsssAs', 'MAXsssdAs', 'MAXsssssAs', 'MAXsSeH', 'MAXdSe', 'MAXssSe', 'MAXaaSe', 'MAXdssSe', 'MAXddssSe', 'MAXsBr', 'MAXsSnH3', 'MAXssSnH2', 'MAXsssSnH', 'MAXssssSn', 'MAXsI', 'MAXsPbH3', 'MAXssPbH2', 'MAXsssPbH', 'MAXssssPb', 'MINsLi', 'MINssBe', 'MINssssBe', 'MINssBH', 'MINsssB', 'MINssssB', 'MINsCH3', 'MINdCH2', 'MINssCH2', 'MINtCH', 'MINdsCH', 'MINaaCH', 'MINsssCH', 'MINddC', 'MINtsC', 'MINdssC', 'MINaasC', 'MINaaaC', 'MINssssC', 'MINsNH3', 'MINsNH2', 'MINssNH2', 'MINdNH', 'MINssNH', 'MINaaNH', 'MINtN', 'MINsssNH', 'MINdsN', 'MINaaN', 'MINsssN', 'MINddsN', 'MINaasN', 'MINssssN', 'MINsOH', 'MINdO', 'MINssO', 'MINaaO', 'MINsF', 'MINsSiH3', 'MINssSiH2', 'MINsssSiH', 'MINssssSi', 'MINsPH2', 'MINssPH', 'MINsssP', 'MINdsssP', 'MINsssssP', 'MINsSH', 'MINdS', 'MINssS', 'MINaaS', 'MINdssS', 'MINddssS', 'MINsCl', 'MINsGeH3', 'MINssGeH2', 'MINsssGeH', 'MINssssGe', 'MINsAsH2', 'MINssAsH', 'MINsssAs', 'MINsssdAs', 'MINsssssAs', 'MINsSeH', 'MINdSe', 'MINssSe', 'MINaaSe', 'MINdssSe', 'MINddssSe', 'MINsBr', 'MINsSnH3', 'MINssSnH2', 'MINsssSnH', 'MINssssSn', 'MINsI', 'MINsPbH3', 'MINssPbH2', 'MINsssPbH', 'MINssssPb'] The limit is marked to the 15.0 %

[+] fitting

[+] transforming

Size of the database, postimputation: (925, 3561)

C-001 C-002 C-003 C-004 C-005 C-006 C-007 ... SsI SsPbH3 SssPbH2 SsssPbH SssssPb SLogP SMR

0 1.0 0.0 0.0 0.0 0.0 4.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 4.19490 88.9710

1 2.0 0.0 0.0 0.0 0.0 2.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.56360 51.4885

2 1.0 1.0 0.0 0.0 1.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.85510 67.2384

3 1.0 1.0 0.0 1.0 3.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 1.77390 78.3157

4 1.0 0.0 0.0 0.0 0.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 2.54542 71.0375

.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...

920 2.0 7.0 3.0 2.0 0.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 4.11138 93.1328

921 0.0 1.0 0.0 0.0 1.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 3.37120 83.0670

922 0.0 0.0 0.0 0.0 2.0 0.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 0.06890 45.3086

923 0.0 5.0 0.0 0.0 0.0 1.0 0.0 ... 0.00000 0.0 0.0 0.0 0.0 2.78510 54.5578

924 0.0 0.0 0.0 0.0 0.0 1.0 0.0 ... 1.68662 0.0 0.0 0.0 0.0 -1.27010 65.9383

[925 rows x 3561 columns]

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-calculated\_imputed\_ytransformed.csv

Do you want to perform any other step?(y/n):

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 2

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/" folder is needed

This file must be called: "BSEP\_NO3D\_local-calculated\_imputed\_ytransformed.csv"

Continue (Y/n)?

[1] Initial feature reduction: infinite, correlated, constant and empty values

0 infinite values

0 features with greater than 0.00 missing values.

1466 features with a correlation magnitude greater than 0.90.

1160 features with a single unique value.

Data has not been one-hot encoded

Removed 2626 features including one-hot features.

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_NO3D\_local-initial\_reduction.csv

Do you want to perform any other step?(y/n): n

Thanks for using NEO!

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME

Models/BSEP/BSEP\_balanced/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local\_final

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 3

Please select your type of model:

[1] Regression

[2] Classification

Your choice (1/2)?: 2

Please input your desired TEST SIZE (enter to: "0.25"): 0.3

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder is needed

This file must be called: "BSEP\_NO3D\_local\_final-initial\_reduction.csv"

Continue (Y/n)?

Traceback (most recent call last):

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1418, in <module>

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 665, in main

initial\_red = read\_dataframe(INPUT\_FILE)

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 97, in read\_dataframe

dataset = pd.read\_csv(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\util\\_decorators.py", line 311, in wrapper

return func(\*args, \*\*kwargs)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 586, in read\_csv

return \_read(filepath\_or\_buffer, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 482, in \_read

parser = TextFileReader(filepath\_or\_buffer, \*\*kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 811, in \_\_init\_\_

self.\_engine = self.\_make\_engine(self.engine)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 1040, in \_make\_engine

return mapping[engine](self.f, \*\*self.options) # type: ignore[call-arg]

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\c\_parser\_wrapper.py", line 51, in \_\_init\_\_

self.\_open\_handles(src, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\base\_parser.py", line 222, in \_open\_handles

self.handles = get\_handle(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\common.py", line 701, in get\_handle

handle = open(

FileNotFoundError: [Errno 2] No such file or directory: 'C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-initial\_reduction.csv'

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local\_final

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 3

Please select your type of model:

[1] Regression

[2] Classification

Your choice (1/2)?: 2

Please input your desired TEST SIZE (enter to: "0.25"): 0.3

A file located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder is needed

This file must be called: "BSEP\_NO3D\_local\_final-initial\_reduction.csv"

Continue (Y/n)?

[+] Generation of train and test sets based in kmeans

OPTIMAL NUMBER OF CLUSTERS: 5

NUMBER OF CLUSTERS: 5

SETS: {2, 38, 8, 75, 183}

ALERTS!!

compound number: 280

SMILE: CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1c3ccccc3c3c4c(c5c6ccccc6n2c5c31)C(=O)NC4 1

compound number: 300

SMILE: CO[C@H]1/C=C/O[C@@]2(C)Oc3c(C)c(O)c4c(O)c(c(/C=N/N5CCN(C6CCCC6)CC5)c(O)c4c3C2=O)NC(=O)/C(C)=C\C=C\[C@H](C)[C@H](O)[C@@H](C)[C@@H](O)[C@@H](C)[C@H](OC(C)=O)[C@@H]1C 1

[280, 300]

you have some molecular alerts. It means that these molecules are quite dissimilar

You can (1) eliminate them or (2) maintain them

What is your choice (1/2)?1

Ok,eliminating

(306, 938)

(304, 938)

NUMBER OF CLUSTERS: 5

SETS: {33, 6, 10, 76, 179}

ALERTS!!

[]

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

0 0 NC(=O)c1cnccn1 0 0.0 ... 0.0 0.00000 -0.42450 0

1 1 O=C(O)P(=O)(O)O 0 0.0 ... 0.0 0.00000 -0.15790 0

2 2 O=c1[nH]cc(F)c(=O)[nH]1 0 0.0 ... 0.0 0.00000 -0.79770 0

3 3 CCCC(CCC)C(=O)O 0 2.0 ... 0.0 0.00000 2.28740 0

4 4 CCCc1cc(=O)[nH]c(=S)[nH]1 0 1.0 ... 0.0 0.00000 1.38499 0

.. ... ... .. ... ... ... ... ... ...

301 301 COCCOC(=O)C1=C(C)NC(C)=C(C(=O)OC(C)C)C1c1cccc(... 1 4.0 ... 0.0 0.00000 2.97080 0

302 302 CCOC(=O)[C@@H]1CC(=O)O[C@H](C)C/C=C/C=C/[C@H](... 1 7.0 ... 0.0 -1.49218 2.62490 2

303 303 Cc1c(C(=O)NN2CCCCC2)nn(-c2ccc(Cl)cc2Cl)c1-c1cc... 1 1.0 ... 0.0 0.00000 5.93862 0

304 304 O=C(NO)C1(CS(=O)(=O)c2ccc(Oc3ccc(Cl)cc3)cc2)CC... 1 0.0 ... 0.0 -1.21794 3.20820 0

305 305 C[C@]12CC[C@H]3[C@@H](CCC4=Cc5oncc5C[C@@]43C)[... 1 2.0 ... 0.0 -1.19292 4.11138 3

[304 rows x 939 columns]

0

1

2

3

4

cluster0

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

0 0 NC(=O)c1cnccn1 0 0.0 ... 0.00000 0.00000 -0.42450 0

1 1 O=C(O)P(=O)(O)O 0 0.0 ... 0.00000 0.00000 -0.15790 0

2 2 O=c1[nH]cc(F)c(=O)[nH]1 0 0.0 ... 0.00000 0.00000 -0.79770 0

3 3 CCCC(CCC)C(=O)O 0 2.0 ... 0.00000 0.00000 2.28740 0

4 4 CCCc1cc(=O)[nH]c(=S)[nH]1 0 1.0 ... 0.00000 0.00000 1.38499 0

.. ... ... .. ... ... ... ... ... ...

296 296 O=C(NC1Cc2ccccc2N(C[C@@H](O)CO)C1=O)c1cc2cc(Cl... 1 0.0 ... 1.57243 0.00000 1.86220 0

298 298 C[C@H]1C[C@H]2[C@@H]3CCC4=CC(=O)C=C[C@]4(C)[C@... 1 3.0 ... 0.00000 -4.20413 2.16500 0

301 301 COCCOC(=O)C1=C(C)NC(C)=C(C(=O)OC(C)C)C1c1cccc(... 1 4.0 ... 0.00000 0.00000 2.97080 0

303 303 Cc1c(C(=O)NN2CCCCC2)nn(-c2ccc(Cl)cc2Cl)c1-c1cc... 1 1.0 ... 0.00000 0.00000 5.93862 0

304 304 O=C(NO)C1(CS(=O)(=O)c2ccc(Oc3ccc(Cl)cc3)cc2)CC... 1 0.0 ... 0.00000 -1.21794 3.20820 0

[179 rows x 939 columns]

index SMILES y C-001 C-002 ... SaasC SaaaC SssssC SLogP cluster

12 12 Nc1ccc(C(=O)NCC(=O)O)cc1 0 0.0 0.0 ... 0.94466 0.0 0.0 0.0832 0

270 270 C(=C/c1ccccc1)\CN1CCN(C(c2ccccc2)c2ccccc2)CC1 1 0.0 0.0 ... 4.03389 0.0 0.0 5.1070 0

56 56 O=C(O)CCC(=O)Nc1ccc(S(=O)(=O)Nc2nccs2)cc1 0 0.0 2.0 ... 0.66747 0.0 0.0 1.7472 0

3 3 CCCC(CCC)C(=O)O 0 2.0 4.0 ... 0.00000 0.0 0.0 2.2874 0

147 147 Nc1nc(=O)c2c([nH]1)NCC(CNc1ccc(C(=O)N[C@@H](CC... 0 0.0 2.0 ... 1.08986 0.0 0.0 -0.7311 0

[5 rows x 939 columns]

cluster1

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

98 98 C[C@@H]1O[C@@H](O[C@H]2C[C@@H](O)[C@]3(CO)[C@H... 0 2.0 ... 0.00000 -5.26775 -1.51500 1

100 100 COc1cc([C@@H]2c3cc4c(cc3[C@@H](O[C@@H]3O[C@@H]... 0 1.0 ... 0.00000 0.00000 1.33860 1

107 107 COc1cc([C@@H]2c3cc4c(cc3[C@@H](O[C@@H]3O[C@@H]... 0 1.0 ... 0.00000 0.00000 1.10450 1

122 122 NCCCC[C@H](NC(=O)[C@@H]1CCCN1C(=O)[C@@H]1CSSC[... 0 0.0 ... 0.00000 0.00000 -6.98050 1

169 169 CC[C@H]1OC(=O)[C@H](C)C(=O)[C@H](C)[C@@H](O[C@... 1 8.0 ... 0.00000 -2.66841 4.92920 1

185 185 CO[C@H]1/C=C/O[C@@]2(C)Oc3c(C)c(O)c4c(O)c(cc(O... 1 8.0 ... -0.47797 -1.98114 4.75412 1

190 190 CC(=O)N[C@H](Cc1ccc2ccccc2c1)C(=O)N[C@H](Cc1cc... 1 4.0 ... 1.83676 0.00000 -0.50613 1

214 214 COC(=O)[C@H]1[C@H]2C[C@@H]3c4[nH]c5cc(OC)ccc5c... 1 0.0 ... 2.26787 0.00000 4.17110 1

242 242 C/C=C/C[C@@H](C)[C@@H](O)[C@H]1C(=O)N[C@@H](CC... 1 17.0 ... 0.00000 0.00000 2.51150 1

256 256 CCCCC(=O)OCC(=O)[C@]1(O)Cc2c(O)c3c(c(O)c2[C@@H... 1 2.0 ... 0.00000 -7.77990 2.46220 1

[10 rows x 939 columns]

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

185 185 CO[C@H]1/C=C/O[C@@]2(C)Oc3c(C)c(O)c4c(O)c(cc(O... 1 8.0 ... -0.47797 -1.98114 4.75412 1

122 122 NCCCC[C@H](NC(=O)[C@@H]1CCCN1C(=O)[C@@H]1CSSC[... 0 0.0 ... 0.00000 0.00000 -6.98050 1

100 100 COc1cc([C@@H]2c3cc4c(cc3[C@@H](O[C@@H]3O[C@@H]... 0 1.0 ... 0.00000 0.00000 1.33860 1

214 214 COC(=O)[C@H]1[C@H]2C[C@@H]3c4[nH]c5cc(OC)ccc5c... 1 0.0 ... 2.26787 0.00000 4.17110 1

256 256 CCCCC(=O)OCC(=O)[C@]1(O)Cc2c(O)c3c(c(O)c2[C@@H... 1 2.0 ... 0.00000 -7.77990 2.46220 1

[5 rows x 939 columns]

cluster2

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

87 87 CC[C@H]1CN2CCc3cc(OC)c(OC)cc3[C@@H]2C[C@@H]1C[... 0 1.0 ... 0.00000 0.00000 4.94340 2

91 91 C=CC[N+]1([C@H]2C[C@H]3[C@@H]4CC[C@H]5C[C@H](O... 0 3.0 ... 0.00000 0.34324 4.40750 2

92 92 CC(C)C(=O)OCC(=O)[C@@]12O[C@H](C3CCCCC3)O[C@@H... 0 4.0 ... 0.00000 -2.29292 4.70390 2

93 93 COc1cccc2c1C(=O)c1c(O)c3c(c(O)c1C2=O)C[C@@](O)... 0 1.0 ... 0.00000 -2.24074 0.00130 2

94 94 COc1cccc2c1C(=O)c1c(O)c3c(c(O)c1C2=O)C[C@@](O)... 0 1.0 ... 0.00000 -2.24074 0.00130 2

95 95 CC(=O)O[C@H]1C[C@@H]2CC[C@@H]3[C@H](CC[C@@]4(C... 0 4.0 ... 0.00000 0.37723 5.96590 2

97 97 CC(=O)O[C@H]1C[C@@H]2CC[C@@H]3[C@H](CC[C@@]4(C... 0 4.0 ... 0.00000 0.37703 6.11050 2

99 99 CN(C)c1cc(NC(=O)CNC(C)(C)C)c(O)c2c1C[C@H]1C[C@... 0 3.0 ... 0.00000 -3.07383 0.51320 2

106 106 C[C@@H]1C/C=C/C=C/C=C/C=C/[C@H](O[C@@H]2O[C@H]... 0 2.0 ... 0.00000 -2.09951 0.11970 2

113 113 CC[C@H]1OC(=O)[C@H](C)[C@@H](O[C@H]2C[C@@](C)(... 0 9.0 ... 0.00000 -4.22646 1.95570 2

114 114 N=C(N)NCCCC[C@@H]1NC(=O)CCSSC[C@@H](C(N)=O)NC(... 0 0.0 ... 1.52264 0.00000 -1.84473 2

116 116 CO[C@@H]1[C@@H](O[C@@H]2O[C@H](C)[C@@H](O[C@@H... 0 6.0 ... 0.00000 -1.47034 2.32510 2

117 117 CCOc1ccc(C[C@H]2NC(=O)CCSSC[C@@H](C(=O)N3CCC[C... 0 4.0 ... 0.00000 0.00000 -3.04460 2

118 118 CC[C@H](C)[C@@H]1NC(=O)[C@H](Cc2ccc(O)cc2)NC(=... 0 4.0 ... 0.00000 0.00000 -3.60860 2

119 119 N=C(N)NCCC[C@@H](NC(=O)[C@@H]1CCCN1C(=O)[C@@H]... 0 0.0 ... 0.00000 0.00000 -4.13203 2

120 120 N=C(N)NCCC[C@H](NC(=O)[C@@H]1CCCN1C(=O)[C@@H]1... 0 0.0 ... 0.00000 0.00000 -5.19483 2

121 121 CCNC(=O)[C@@H]1CCCN1C(=O)[C@H](CCCNC(=N)N)NC(=... 0 5.0 ... 1.43055 0.00000 -1.22843 2

123 123 CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@H... 0 5.0 ... 1.39300 -0.91279 -2.89603 2

154 154 Cn1c(=O)n(-c2ccc(C(C)(C)C#N)cc2)c2c3cc(-c4cnc5... 1 2.0 ... 5.26984 -0.60947 5.89378 2

157 157 C/C=C/C[C@@H](C)[C@@H](O)[C@H]1C(=O)N[C@@H](CC... 1 19.0 ... 0.00000 0.00000 3.65750 2

173 173 CCCc1nc2c(C)cc(-c3nc4ccccc4n3C)cc2n1Cc1ccc(-c2... 1 2.0 ... 4.20950 0.00000 7.26442 2

176 176 CC(=O)CC[C@H]1C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](... 1 5.0 ... 0.00000 0.00000 3.02270 2

177 177 CC1OC(=O)C(C(C)C)NC(=O)C(C(C)C)OC(=O)C(C(C)C)N... 1 21.0 ... 0.00000 0.00000 2.34330 2

192 192 C=CC[C@@H]1/C=C(\C)C[C@H](C)C[C@H](OC)[C@H]2O[... 1 5.0 ... 0.00000 -2.51025 4.63900 2

198 198 C[C@]12CC[C@H]3[C@H]([C@@H]1[C@@H]1C[C@@H]1[C@... 1 2.0 ... 0.00000 0.32094 4.30590 2

213 213 CS(=O)(=O)CCNCc1ccc(-c2ccc3ncnc(Nc4ccc(OCc5ccc... 1 0.0 ... 1.53092 0.00000 6.13910 2

222 222 C/C=C/CC(C)C(O)C1C(=O)NC(CC)C(=O)N(C)C(C)C(=O)... 1 19.0 ... 0.00000 0.00000 3.65750 2

246 246 CC(C)[C@@H]1NC(=O)[C@H](C)OC(=O)[C@@H](C(C)C)N... 1 21.0 ... 0.00000 0.00000 2.34330 2

253 253 C[C@@H](O)[C@@H]1NC(=O)[C@H](CCCCN)NC(=O)[C@@H... 1 2.0 ... 1.53024 0.00000 -0.80540 2

254 254 CC[C@@H]1/C=C(\C)C[C@H](C)C[C@H](OC)[C@H]2O[C@... 1 6.0 ... 0.00000 -2.49150 5.71940 2

288 288 C/C=C/CC(C)C(O)C1C(=O)NC(CC)C(=O)N(C)CC(=O)N(C... 1 17.0 ... 0.00000 0.00000 3.26900 2

292 292 C/C=C/C[C@@H](C)[C@@H](O)[C@H]1C(=O)N[C@@H](CC... 1 17.0 ... 0.00000 0.00000 3.26900 2

302 302 CCOC(=O)[C@@H]1CC(=O)O[C@H](C)C/C=C/C=C/[C@H](... 1 7.0 ... 0.00000 -1.49218 2.62490 2

[33 rows x 939 columns]

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

176 176 CC(=O)CC[C@H]1C(=O)N[C@@H](C(C)C)C(=O)N[C@@H](... 1 5.0 ... 0.0 0.00000 3.02270 2

113 113 CC[C@H]1OC(=O)[C@H](C)[C@@H](O[C@H]2C[C@@](C)(... 0 9.0 ... 0.0 -4.22646 1.95570 2

120 120 N=C(N)NCCC[C@H](NC(=O)[C@@H]1CCCN1C(=O)[C@@H]1... 0 0.0 ... 0.0 0.00000 -5.19483 2

94 94 COc1cccc2c1C(=O)c1c(O)c3c(c(O)c1C2=O)C[C@@](O)... 0 1.0 ... 0.0 -2.24074 0.00130 2

292 292 C/C=C/C[C@@H](C)[C@@H](O)[C@H]1C(=O)N[C@@H](CC... 1 17.0 ... 0.0 0.00000 3.26900 2

[5 rows x 939 columns]

cluster3

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

63 63 C=CCN1CC[C@]23c4c5ccc(O)c4O[C@H]2C(=O)CC[C@@]3... 0 0.0 ... 0.00000 -1.71706 1.30140 3

64 64 C=C1CC[C@@]2(O)[C@H]3Cc4ccc(O)c5c4[C@@]2(CCN3C... 0 0.0 ... 0.00000 -1.20164 2.51250 3

65 65 O=C1CC[C@@]2(O)[C@H]3Cc4ccc(O)c5c4[C@@]2(CCN3C... 0 0.0 ... 0.00000 -1.61164 1.52540 3

66 66 CC[C@@]1(O)C(=O)OCc2c1cc1n(c2=O)Cc2cc3ccccc3nc2-1 0 1.0 ... 1.85692 -1.78975 2.07960 3

67 67 C[C@]12CC[C@H]3[C@@H](CC=C4C[C@@H](O)CC[C@@]43... 0 2.0 ... 0.00000 0.63950 5.39860 3

.. ... ... .. ... ... ... ... ... ...

285 285 CC1(C)OC(=O)N([C@H]2CC[C@H](c3cnc(N)c(-c4ncccn... 1 2.0 ... 0.00000 -0.56870 5.11910 3

294 294 CC(C)(C)NC(=O)[C@@H]1C[C@@H]2CCCC[C@@H]2CN1C[C... 1 3.0 ... 1.47028 -0.40591 3.09240 3

297 297 COc1ccc2c(c1)C(=O)N(CCc1ccc(S(=O)(=O)NC(=O)NC3... 1 2.0 ... 0.00000 -0.87766 3.51850 3

299 299 CCC(=O)O[C@]1(C(=O)CCl)[C@@H](C)C[C@H]2[C@@H]3... 1 4.0 ... 0.00000 -5.54011 4.10310 3

305 305 C[C@]12CC[C@H]3[C@@H](CCC4=Cc5oncc5C[C@@]43C)[... 1 2.0 ... 0.00000 -1.19292 4.11138 3

[76 rows x 939 columns]

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

90 90 Cc1nnc(C(C)C)n1[C@H]1C[C@@H]2CC[C@H](C1)N2CC[C... 0 3.0 ... 0.0 -2.62276 5.95092 3

215 215 C=C(C)[C@H]1Cc2c(ccc3c2O[C@@H]2COc4cc(OC)c(OC)... 1 1.0 ... 0.0 0.00000 3.70330 3

233 233 COc1cc(-c2ccc(C3CCc4ccc([C@H](C5CC5)[C@H](C)C(... 1 1.0 ... 0.0 0.00000 5.57190 3

63 63 C=CCN1CC[C@]23c4c5ccc(O)c4O[C@H]2C(=O)CC[C@@]3... 0 0.0 ... 0.0 -1.71706 1.30140 3

297 297 COc1ccc2c(c1)C(=O)N(CCc1ccc(S(=O)(=O)NC(=O)NC3... 1 2.0 ... 0.0 -0.87766 3.51850 3

[5 rows x 939 columns]

cluster4

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

207 207 CO[C@H]1C[C@@H]2CC[C@@H](C)[C@@](O)(O2)C(=O)C(... 1 7.0 ... 0.0000 -2.43345 6.19720 4

210 210 CC[C@@H]([C@H](C)O)n1ncn(-c2ccc(N3CCN(c4ccc(OC... 1 2.0 ... 0.0000 -1.01860 4.57320 4

219 219 CC(=O)O[C@H]1C(=O)[C@]2(C)[C@@H](O)C[C@H]3OC[C... 1 6.0 ... 0.0000 -7.87688 3.73570 4

241 241 CO[C@H]1/C=C/O[C@@]2(C)Oc3c(C)c(O)c4c(O)c(c(/C... 1 8.0 ... -0.5529 -2.04257 4.33542 4

247 247 CCC(C)n1ncn(-c2ccc(N3CCN(c4ccc(OCC5COC(Cn6cncn... 1 2.0 ... 0.0000 -1.15850 5.57730 4

250 250 CO[C@H]1C[C@@H]2CC[C@@H](C)[C@@](O)(O2)C(=O)C(... 1 8.0 ... 0.0000 -3.95641 5.72240 4

[6 rows x 939 columns]

index SMILES y C-001 ... SaaaC SssssC SLogP cluster

210 210 CC[C@@H]([C@H](C)O)n1ncn(-c2ccc(N3CCN(c4ccc(OC... 1 2.0 ... 0.0 -1.01860 4.5732 4

250 250 CO[C@H]1C[C@@H]2CC[C@@H](C)[C@@](O)(O2)C(=O)C(... 1 8.0 ... 0.0 -3.95641 5.7224 4

219 219 CC(=O)O[C@H]1C(=O)[C@]2(C)[C@@H](O)C[C@H]3OC[C... 1 6.0 ... 0.0 -7.87688 3.7357 4

207 207 CO[C@H]1C[C@@H]2CC[C@@H](C)[C@@](O)(O2)C(=O)C(... 1 7.0 ... 0.0 -2.43345 6.1972 4

[4 rows x 939 columns]

index SMILES y C-001 C-002 ... SaasC SaaaC SssssC SLogP cluster

12 12 Nc1ccc(C(=O)NCC(=O)O)cc1 0 0.0 0.0 ... 0.94466 0.0 0.0 0.0832 0

270 270 C(=C/c1ccccc1)\CN1CCN(C(c2ccccc2)c2ccccc2)CC1 1 0.0 0.0 ... 4.03389 0.0 0.0 5.1070 0

56 56 O=C(O)CCC(=O)Nc1ccc(S(=O)(=O)Nc2nccs2)cc1 0 0.0 2.0 ... 0.66747 0.0 0.0 1.7472 0

3 3 CCCC(CCC)C(=O)O 0 2.0 4.0 ... 0.00000 0.0 0.0 2.2874 0

147 147 Nc1nc(=O)c2c([nH]1)NCC(CNc1ccc(C(=O)N[C@@H](CC... 0 0.0 2.0 ... 1.08986 0.0 0.0 -0.7311 0

[5 rows x 939 columns]

Train set contains:

108 negative values

104 positive values

ratio neg / pos: 1.0384615384615385

Test set contains:

46 negative values

46 positive values

ratio neg / pos: 1.0

If you find this imbalanced, try to decomment line 44 of split\_by\_kmeans.py module. It can give an error!

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-cleaned\_from\_kmeans.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_set.csv

Do you want to perform any other step?(y/n): y

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 4

[+] Descriptor standarization

Please select the method to standarize the descriptors:

[1] StandardScaler

[2] MinMaxScaler

Your choice (1/2)?: 1

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-train\_set.csv"

"BSEP\_NO3D\_local\_final-test\_set.csv"

Continue (Y/n)?

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-stand\_train\_set.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-stand\_test\_set.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-alldataset.sca

Do you want to perform any other step?(y/n):

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 6

[+] Reduction by FI based on lgbm

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?y

Please define if your model is for [1] classification or [2] regression:1

Please define your parameters for lgbm selection for classification parameters:

eval\_metric (l2/auc/binary\_logloss):

Training Gradient Boosting Model

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[93] valid\_0's binary\_logloss: 0.3748

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[42] valid\_0's binary\_logloss: 0.441151

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[64] valid\_0's binary\_logloss: 0.501908

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[89] valid\_0's binary\_logloss: 0.402745

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[58] valid\_0's binary\_logloss: 0.484669

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[40] valid\_0's binary\_logloss: 0.468414

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[59] valid\_0's binary\_logloss: 0.396554

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[83] valid\_0's binary\_logloss: 0.49645

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[50] valid\_0's binary\_logloss: 0.447678

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[78] valid\_0's binary\_logloss: 0.358568

530 features with zero importance after one-hot encoding.

Please, close the plots to continue...

215 features required for 0.90 of cumulative importance

Please, close the plots to continue...

214 features required for cumulative importance of 0.90 after one hot encoding.

721 features do not contribute to cumulative importance of 0.90.

feature importance normalized\_importance cumulative\_importance

0 SLogP 50.7 0.124539 0.124539

1 ATS0Z 12.7 0.031196 0.155736

2 SMR\_VSA10 9.0 0.022108 0.177843

3 EState\_VSA3 8.7 0.021371 0.199214

4 PW2 7.5 0.018423 0.217637

.. ... ... ... ...

572 NaaO 0.0 0.000000 1.000000

571 NssS 0.0 0.000000 1.000000

570 NaaS 0.0 0.000000 1.000000

584 C-011 0.0 0.000000 1.000000

934 B04[N-I] 0.0 0.000000 1.000000

[935 rows x 4 columns]

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_featured\_importances.csv

Now you can select a number of features.

Please, indicate the number of selected features: 25

Selected features:

['SLogP', 'ATS0Z', 'SMR\_VSA10', 'EState\_VSA3', 'PW2', 'X3solA', 'AATS8i', 'GATS7dv', 'AATS7i', 'SssCH2', 'AATS8s', 'GATS2i', 'SIC3', 'JGI10', 'EState\_VSA6', 'AATS4p', 'GATS6d', 'TI2', 'SaasC', 'AATS3s', 'AATS5i', 'BELZ0', 'AATS5dv', 'ATSC3d', 'X2A']

Traceback (most recent call last):

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1418, in <module>

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 710, in main

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 805, in main

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1072, in main

full\_reimputed\_rescaled = re\_imputation\_scalation(PATH,MODEL,selected\_features)

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 264, in re\_imputation\_scalation

full\_dataset = read\_dataframe(INPUT\_FILE)

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 97, in read\_dataframe

dataset = pd.read\_csv(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\util\\_decorators.py", line 311, in wrapper

return func(\*args, \*\*kwargs)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 586, in read\_csv

return \_read(filepath\_or\_buffer, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 482, in \_read

parser = TextFileReader(filepath\_or\_buffer, \*\*kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 811, in \_\_init\_\_

self.\_engine = self.\_make\_engine(self.engine)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 1040, in \_make\_engine

return mapping[engine](self.f, \*\*self.options) # type: ignore[call-arg]

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\c\_parser\_wrapper.py", line 51, in \_\_init\_\_

self.\_open\_handles(src, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\base\_parser.py", line 222, in \_open\_handles

self.handles = get\_handle(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\common.py", line 701, in get\_handle

handle = open(

FileNotFoundError: [Errno 2] No such file or directory: 'C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-calculated\_preimputation.csv'

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local\_final

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 6

[+] Reduction by FI based on lgbm

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?

Please define if your model is for [1] classification or [2] regression:1

Please define your parameters for lgbm selection for classification parameters:

eval\_metric (l2/auc/binary\_logloss):

Training Gradient Boosting Model

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[93] valid\_0's binary\_logloss: 0.3748

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[42] valid\_0's binary\_logloss: 0.441151

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[64] valid\_0's binary\_logloss: 0.501908

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[89] valid\_0's binary\_logloss: 0.402745

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[58] valid\_0's binary\_logloss: 0.484669

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[40] valid\_0's binary\_logloss: 0.468414

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[59] valid\_0's binary\_logloss: 0.396554

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[83] valid\_0's binary\_logloss: 0.49645

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[50] valid\_0's binary\_logloss: 0.447678

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[78] valid\_0's binary\_logloss: 0.358568

530 features with zero importance after one-hot encoding.

Please, close the plots to continue...

215 features required for 0.90 of cumulative importance

Please, close the plots to continue...

214 features required for cumulative importance of 0.90 after one hot encoding.

721 features do not contribute to cumulative importance of 0.90.

feature importance normalized\_importance cumulative\_importance

0 SLogP 50.7 0.124539 0.124539

1 ATS0Z 12.7 0.031196 0.155736

2 SMR\_VSA10 9.0 0.022108 0.177843

3 EState\_VSA3 8.7 0.021371 0.199214

4 PW2 7.5 0.018423 0.217637

.. ... ... ... ...

572 NaaO 0.0 0.000000 1.000000

571 NssS 0.0 0.000000 1.000000

570 NaaS 0.0 0.000000 1.000000

584 C-011 0.0 0.000000 1.000000

934 B04[N-I] 0.0 0.000000 1.000000

[935 rows x 4 columns]

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_featured\_importances.csv

Now you can select a number of features.

Please, indicate the number of selected features: 25

Selected features:

['SLogP', 'ATS0Z', 'SMR\_VSA10', 'EState\_VSA3', 'PW2', 'X3solA', 'AATS8i', 'GATS7dv', 'AATS7i', 'SssCH2', 'AATS8s', 'GATS2i', 'SIC3', 'JGI10', 'EState\_VSA6', 'AATS4p', 'GATS6d', 'TI2', 'SaasC', 'AATS3s', 'AATS5i', 'BELZ0', 'AATS5dv', 'ATSC3d', 'X2A']

Traceback (most recent call last):

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1418, in <module>

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1072, in main

full\_reimputed\_rescaled = re\_imputation\_scalation(PATH,MODEL,selected\_features)

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 264, in re\_imputation\_scalation

full\_dataset = read\_dataframe(INPUT\_FILE)

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 97, in read\_dataframe

dataset = pd.read\_csv(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\util\\_decorators.py", line 311, in wrapper

return func(\*args, \*\*kwargs)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 586, in read\_csv

return \_read(filepath\_or\_buffer, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 482, in \_read

parser = TextFileReader(filepath\_or\_buffer, \*\*kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 811, in \_\_init\_\_

self.\_engine = self.\_make\_engine(self.engine)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\readers.py", line 1040, in \_make\_engine

return mapping[engine](self.f, \*\*self.options) # type: ignore[call-arg]

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\c\_parser\_wrapper.py", line 51, in \_\_init\_\_

self.\_open\_handles(src, kwds)

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\parsers\base\_parser.py", line 222, in \_open\_handles

self.handles = get\_handle(

File "C:\Users\Enrique\miniconda3\envs\Protocosas\lib\site-packages\pandas\io\common.py", line 701, in get\_handle

handle = open(

FileNotFoundError: [Errno 2] No such file or directory: 'C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-calculated\_preimputation.csv'

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local\_final

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 6

[+] Reduction by FI based on lgbm

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?

Please define if your model is for [1] classification or [2] regression:1

Please define your parameters for lgbm selection for classification parameters:

eval\_metric (l2/auc/binary\_logloss):

Training Gradient Boosting Model

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[93] valid\_0's binary\_logloss: 0.3748

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[42] valid\_0's binary\_logloss: 0.441151

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[64] valid\_0's binary\_logloss: 0.501908

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[89] valid\_0's binary\_logloss: 0.402745

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[58] valid\_0's binary\_logloss: 0.484669

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[40] valid\_0's binary\_logloss: 0.468414

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[59] valid\_0's binary\_logloss: 0.396554

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[83] valid\_0's binary\_logloss: 0.49645

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[50] valid\_0's binary\_logloss: 0.447678

Training until validation scores don't improve for 100 rounds

Early stopping, best iteration is:

[78] valid\_0's binary\_logloss: 0.358568

530 features with zero importance after one-hot encoding.

Please, close the plots to continue...

215 features required for 0.90 of cumulative importance

Please, close the plots to continue...

214 features required for cumulative importance of 0.90 after one hot encoding.

721 features do not contribute to cumulative importance of 0.90.

feature importance normalized\_importance cumulative\_importance

0 SLogP 50.7 0.124539 0.124539

1 ATS0Z 12.7 0.031196 0.155736

2 SMR\_VSA10 9.0 0.022108 0.177843

3 EState\_VSA3 8.7 0.021371 0.199214

4 PW2 7.5 0.018423 0.217637

.. ... ... ... ...

572 NaaO 0.0 0.000000 1.000000

571 NssS 0.0 0.000000 1.000000

570 NaaS 0.0 0.000000 1.000000

584 C-011 0.0 0.000000 1.000000

934 B04[N-I] 0.0 0.000000 1.000000

[935 rows x 4 columns]

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_featured\_importances.csv

Now you can select a number of features.

Please, indicate the number of selected features: 25

Selected features:

['SLogP', 'ATS0Z', 'SMR\_VSA10', 'EState\_VSA3', 'PW2', 'X3solA', 'AATS8i', 'GATS7dv', 'AATS7i', 'SssCH2', 'AATS8s', 'GATS2i', 'SIC3', 'JGI10', 'EState\_VSA6', 'AATS4p', 'GATS6d', 'TI2', 'SaasC', 'AATS3s', 'AATS5i', 'BELZ0', 'AATS5dv', 'ATSC3d', 'X2A']

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_GBM.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_GBM.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features?(y/n): n

Please, indicate the number of selected features:

Traceback (most recent call last):

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1418, in <module>

main()

File "C:\Users\Enrique\Documents\generate\_models\NEO\NEO.py", line 1064, in main

SELECT\_K\_VARS = int(input('Please, indicate the number of selected features: '))

KeyboardInterrupt

^C

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>

(Protocosas) C:\Users\Enrique\Documents\generate\_models\NEO>python NEO.py

#########################################################################

######################### WELCOME TO NEO script #########################

#########################################################################

This script will allow you to:

- eliminate 3D descriptors

- "y" transformation

- perform the initial unsupervised feature reduction

- perform the train/test split based on kmeans

- descriptor standarization

- select the relevant features based on:

· Recursive feature elimination (RFE)

· Feature importance (FI) based on Ligth gradient boosting machine (LGBM)

· Permutation importance (PI)

- select your own features features

Please input your PATH (enter to: "../data/Af\_MIC80\_definitva/no3D/OWNdesc/"): C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/

Please input your MODEL NAME (enter to: Af\_MIC80\_no3D): BSEP\_NO3D\_local\_final

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 5

[+] Reduction by RFE

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?

Please select your estimator for perform RFE:

For regression models:

[1] SVR with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[2] Ridge

[3] LinearRegression

For classification models:

[4] SVC with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[5] LogisticRegression

[6] LDA

Your choice: 4

Now you can select a number of features and steps.

Please, indicate the number of selected features: 25

Please, indicate the number of steps:1

Please wait...

Selected features:

['C-022', 'N-079', 'AATS1i', 'ATSC6se', 'MATS1se', 'X0vA', 'nMultiple', 'nR\_3\_True\_True\_False\_True', 'nR\_11\_True\_True\_True\_True', 'JGI10', 'F04[C-Cl]', 'B05[S-Cl]', 'B07[N-N]', 'B07[N-Cl]', 'nArCOO', 'nDihydPyr', 'nEster', 'nArNitro', 'PEOE\_VSA4', 'PEOE\_VSA6', 'SLogP\_VSA4', 'EState\_VSA3', 'EState\_VSA8', 'NaasC', 'SLogP']

R\_squared 0.9811320754716981

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): y

Do you want to perform any other step?(y/n):

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 5

[+] Reduction by RFE

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?4

Incorrect input. Continue (Y/n)?

Incorrect input. Continue (Y/n)?y

Please select your estimator for perform RFE:

For regression models:

[1] SVR with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[2] Ridge

[3] LinearRegression

For classification models:

[4] SVC with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[5] LogisticRegression

[6] LDA

Your choice: 4

Now you can select a number of features and steps.

Please, indicate the number of selected features: 25

Please, indicate the number of steps:2

Please wait...

Selected features:

['C-024', 'H-046', 'H-048', 'AATS1i', 'ATSC6d', 'nR\_7\_True\_False\_False\_False', 'nR\_10\_True\_True\_None\_None', 'nR\_11\_True\_True\_False\_False', 'JGI10', 'B03[C-O]', 'F04[C-Cl]', 'B05[S-Cl]', 'B06[O-O]', 'B10[N-Cl]', 'nArCOO', 'nHOCCN', 'nEster', 'nLactone', 'nQuatN', 'EState\_VSA3', 'EState\_VSA6', 'EState\_VSA7', 'EState\_VSA8', 'NaasC', 'SLogP']

R\_squared 0.9858490566037735

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n):

Please, indicate the number of selected features: 25

Please, indicate the number of steps:1

Please wait...

Selected features:

['C-022', 'N-079', 'AATS1i', 'ATSC6se', 'MATS1se', 'X0vA', 'nMultiple', 'nR\_3\_True\_True\_False\_True', 'nR\_11\_True\_True\_True\_True', 'JGI10', 'F04[C-Cl]', 'B05[S-Cl]', 'B07[N-N]', 'B07[N-Cl]', 'nArCOO', 'nDihydPyr', 'nEster', 'nArNitro', 'PEOE\_VSA4', 'PEOE\_VSA6', 'SLogP\_VSA4', 'EState\_VSA3', 'EState\_VSA8', 'NaasC', 'SLogP']

R\_squared 0.9811320754716981

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:2

Please wait...

Selected features:

['C-024', 'H-046', 'H-048', 'AATS1i', 'ATSC6d', 'nR\_7\_True\_False\_False\_False', 'nR\_10\_True\_True\_None\_None', 'nR\_11\_True\_True\_False\_False', 'JGI10', 'B03[C-O]', 'F04[C-Cl]', 'B05[S-Cl]', 'B06[O-O]', 'B10[N-Cl]', 'nArCOO', 'nHOCCN', 'nEster', 'nLactone', 'nQuatN', 'EState\_VSA3', 'EState\_VSA6', 'EState\_VSA7', 'EState\_VSA8', 'NaasC', 'SLogP']

R\_squared 0.9858490566037735

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:3

Please wait...

Selected features:

['C-009', 'H-046', 'O-060', 'AATS1i', 'AATS3p', 'ATSC4d', 'ATSC6se', 'AATSC6v', 'BELd-1', 'BELv-1', 'nR\_7\_True\_False\_False\_True', 'nR\_10\_True\_True\_None\_None', 'F04[C-Cl]', 'B07[N-N]', 'F08[N-N]', 'nEster', 'nPiperzine', 'nQuatN', 'PEOE\_VSA4', 'PEOE\_VSA6', 'SLogP\_VSA4', 'EState\_VSA3', 'EState\_VSA7', 'SaasC', 'SLogP']

R\_squared 0.9528301886792453

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): y

Do you want to perform any other step?(y/n): y

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 5

[+] Reduction by RFE

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?5

Incorrect input. Continue (Y/n)?

Incorrect input. Continue (Y/n)?y

Please select your estimator for perform RFE:

For regression models:

[1] SVR with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[2] Ridge

[3] LinearRegression

For classification models:

[4] SVC with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[5] LogisticRegression

[6] LDA

Your choice: 5

Now you can select a number of features and steps.

Please, indicate the number of selected features: 25

Please, indicate the number of steps:1

Please wait...

Selected features:

['C-008', 'C-009', 'AATSC3c', 'AATSC6v', 'nMultiple', 'nR\_3\_True\_True\_False\_True', 'nR\_7\_True\_False\_False\_True', 'nR\_7\_True\_False\_False\_False', 'nR\_11\_True\_True\_True\_True', 'F04[C-Cl]', 'B07[N-Cl]', 'F08[N-N]', 'nArCOO', 'nDihydPyr', 'nEster', 'nLactone', 'nParaHyd', 'nQuatN', 'PEOE\_VSA4', 'PEOE\_VSA6', 'EState\_VSA3', 'EState\_VSA6', 'EState\_VSA7', 'NaasC', 'SLogP']

R\_squared 0.9669811320754716

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:2

Please wait...

Selected features:

['C-024', 'H-046', 'AATS3i', 'GATS6d', 'X4vA', 'nMultiple', 'nR\_3\_True\_True\_True\_False', 'nR\_3\_True\_True\_False\_True', 'nR\_7\_True\_False\_False\_False', 'nR\_11\_True\_True\_True\_True', 'F04[C-Cl]', 'B06[O-O]', 'B07[N-O]', 'B07[N-Cl]', 'nArCOO', 'nEster', 'nLactone', 'nParaHyd', 'nQuatN', 'PEOE\_VSA6', 'EState\_VSA3', 'EState\_VSA8', 'VSA\_EState6', 'NaasC', 'SLogP']

R\_squared 0.9669811320754716

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:3

Please wait...

Selected features:

['H-048', 'AATS0i', 'AATS1i', 'AATSC6dv', 'GATS6d', 'X0vA', 'nMultiple', 'nR\_3\_True\_True\_False\_True', 'nR\_7\_True\_False\_False\_False', 'nR\_11\_True\_True\_True\_True', 'F04[C-Cl]', 'B06[O-O]', 'B07[N-O]', 'B07[N-Cl]', 'nArCOO', 'nEster', 'nLactone', 'nParaHyd', 'nQuatN', 'PEOE\_VSA4', 'PEOE\_VSA6', 'SMR\_VSA7', 'EState\_VSA3', 'EState\_VSA8', 'SLogP']

R\_squared 0.9669811320754716

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): y

Do you want to perform any other step?(y/n): y

######################### MAIN MENU #########################

Please select what do you want to do:

[01] Elimination of 3D descriptors [your dataset will be saved as [Name]\_no3D]

[1] "y" transformation + dataset random order + Knn imputation

[2] Initial feature reduction: infinite, correlated, constant and empty values

[3] Generation of train and test sets based in kmeans

[4] Descriptor standarization

[5] Feature selection by RFE

[6] Feature selection by FI based on LGBM

[7] Feature selection by Permutation importance

[8] Select own features (inside the script)

[0] Exit NEO

Your choice: 5

[+] Reduction by RFE

Two files located in "C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/" folder are needed

These files must be called:

"BSEP\_NO3D\_local\_final-stand\_train\_set.csv"

"BSEP\_NO3D\_local\_final-stand\_test\_set.csv"

Continue (Y/n)?

Please select your estimator for perform RFE:

For regression models:

[1] SVR with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[2] Ridge

[3] LinearRegression

For classification models:

[4] SVC with linear kernel [remember that other kernels can be used:‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’]

[5] LogisticRegression

[6] LDA

Your choice: 6

Now you can select a number of features and steps.

Please, indicate the number of selected features: 25

Please, indicate the number of steps:1

Please wait...

Selected features:

['C-001', 'C-024', 'H-046', 'F-084', 'AATS3p', 'AATS5Z', 'ATSC1se', 'AATSC2dv', 'AATSC6p', 'AATSC6i', 'MATS1s', 'MATS2s', 'MATS5p', 'GATS1s', 'GATS5p', 'nR\_3\_True\_False\_False\_False', 'nR\_11\_True\_True\_True\_True', 'ESpm03d', 'PW2', 'MAXDN', 'B07[S-F]', 'B08[N-O]', 'nEster', 'nHal', 'EState\_VSA2']

R\_squared 0.8867924528301887

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:2

Please wait...

Selected features:

['C-003', 'C-004', 'C-009', 'O-059', 'I-099', 'AATS0se', 'AATS1Z', 'AATS1p', 'AATS1i', 'AATSC1dv', 'AATSC2se', 'AATSC3s', 'MATS3s', 'GATS6s', 'X2A', 'nCL', 'nR\_3\_True\_False\_False\_False', 'nR\_11\_True\_True\_True\_True', 'SIC0', 'B01[N-S]', 'B08[C-Br]', 'F10[C-F]', 'nHal', 'VSA\_EState5', 'SLogP']

R\_squared 0.9056603773584906

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)

Do you agree with that number of features and steps?(y/n): n

Please, indicate the number of selected features: 25

Please, indicate the number of steps:3

Please wait...

Selected features:

['C-009', 'C-024', 'H-046', 'N-068', 'AATS8Z', 'ATSC4v', 'ATSC6p', 'AATSC1v', 'AATSC6v', 'AATSC6p', 'MATS5p', 'MATS6p', 'GATS5dv', 'GATS6i', 'GATS7dv', 'GATS7s', 'X1solA', 'X4A', 'X4solA', 'nCL', 'RBF', 'nR\_3\_True\_False\_False\_False', 'ESpm03d', 'B08[O-I]', 'nArCOO']

R\_squared 0.8915094339622641

Size of the database, preimputation: (212, 25)

[+] fitting

[+] transforming

Size of the database, postimputation: (212, 25)

The following files have been created:

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-test\_reduction\_RFE.csv

C:/Users/Enrique/Documents/New ProtoADME Models/BSEP/BSEP\_balanced/BSEP\_NO3D\_local\_final-train\_set.sca

train\_set (212, 27)

test\_set (92, 27)