Machine-OLF-action: Mapping Odor-Receptor Space

Software User Manual (v0.0.1- DRAFT)

Table of Contents

Software Usage

If a user has some chemical compounds dataset consisting of compound names, compound SMILEs and their activation statuses (0 or 1) on a target (here, 0 indicates that chemical compound will not activate the target, 1 indicates that chemical compound will activate the target), then this software can help the user to create machine learning-based binary classification models and use these models to find new novel chemical compounds and predict their ability to activate the target.

As a couple of example use cases, the software is used

1) For discovery of novel ligands that can bind to the “OR1A1” ola-factory receptor. The training data consisted of ligands curated from the existing literature.
2) For the discovery of novel repellents that can keep mosquitos away. The training data consisted of repellents curated from existing literature.

Software User Guide

The software has two major components, details of which are described in the below sections

1) **Create a new Job** - This component allows users to upload their training data (called dataset) and configure various options available on a page to create binary classification models from the data. It will also allow the user to search similar activating compounds in various compound databases and predict their activation behavior on the target. We will call this page, the job configuration page. All the details are present in the respective section.
2) **View all Jobs** - A user may choose to create multiple jobs based on different datasets s/he may have or s/he may choose to run different experiments using different configurations on the same dataset. A job needs to be created for each of these scenarios. All the jobs which are created by the user are listed here. A user can view detailed results, various intermediated data files, status, and logs related to a job from this page.

Create a new Job

Paste software screenshot of “Create new Job“ page here

URL of the page:
http://127.0.0.1:5748/create-job
Note: Host and port 5748 can change based on the software installation

There are a total of 6 steps with various settings that can help the user to discover novel compounds similar to activating compounds in their training dataset. Details of each of the steps

are mentioned below.

Note: The output dataset of each user selected step or sub-step is used as input of the next user selected step or substep.

Step 1: Choose File

This is a mandatory step where a user needs to upload his/her chemical compound ***dataset*** with activation status on a target. All the classification models are trained and validated on this dataset.

Dataset File Format

The compound dataset should be uploaded as a CSV file with exactly three fields with below-given specifications

1) **Compound Name:** A string indicating the name of the chemical compound.
2) **Compound SMILE:** A string-based SMILE representation of the compound.
3) **Activation Status:** An integer with allowed values as 1 or 0 based on whether the compound will activate a target or not.

It is recommended to have the ratio of activating compounds to non-activating compounds to be at least 15%:85%. The more the ratio is balanced the better are the chances of enhanced classification performance.

Step 2: Feature Generation

Classification based machine learning algorithms can only work with numeric data, so a tool is needed to convert user compound data uploaded in step 1 from string format (SMILEs representation) to the numeric format.

Generate Chemical Descriptors

This sub-step helps just do that, it converts compound data uploaded in step 1 to numeric molecular descriptors based features which can be then used to feed to machine learning algorithms.

Numeric features of a chemical compound, represented using SMILE, can be extracted using open-source tools like PaDEL and Mordred. There are other tools also but the current version software uses python version of PaDEL and Mordred to generate numeric features from the compound's SMILE. In future updates, other molecular descriptor calculators can be integrated.

The user thus must mandatorily choose at least one of the settings (molecule descriptor generator) from the given choices (like PaDEL, Mordred). Of course a user can choose to select more than one descriptor and use them for experimentation and comparison of results with each other.
PaDEL generates 1875 descriptors (1444 1D, 2D descriptors and 431 3D descriptors) which will

be further used by machine learning algorithms. The details of all the features can be found *[give document reference here]*

Mordred is not integrated yet, write description after it's integration.

Choose Preprocessing Steps

The molecular descriptors (numeric features) generated from the above sub-step may have many data issues like missing values of some descriptors for some of the compounds, same value of the descriptor across all compounds or there may be a class imbalance between activation statuses of user uploaded compounds. All these issues need to be rectified before feeding this data to machine learning algorithms for error free training of the prediction model. Preprocessing sub-step helps achieve just that. It can help by removing unwanted features, by imputing missing values, by handling class imbalance so that optimal predictive performance on ML algorithms can be achieved.

Handling Missing value

This step helps users to remove features (using setting "Prune Columns") which have missing values above a particular percentage threshold. Users can enter his/her own threshold value in percentage. It is recommended to keep threshold value as 75%, so that all features which have missing values more than 75% are removed from data.

The columns which have missing values less than the user specified threshold needs to be imputed. User can select to impute missing values using mean of the feature using setting call "Impute with Mean"

Variance Removal

The features which have low variance or zero variance (all the values of the features are the same) should be removed as they don't add any discriminatory information to classification algorithms. This setting allows the user to remove all the features which are having a variance lower than the specified threshold. It is recommended to at least remove the features with 0 variance by keeping the threshold value as 0.

Correlation Check

When two features are correlated, only one of them should be used and the other can be removed as it does not add any additional information gain for the classifier. This setting allows the user to remove one of the correlated features having correlation greater than a particular percentage threshold. It is recommended to remove one of the features out of two features having a correlation of at least 95%.

Normalize

If some feature values are on scale of tens and others are on scale of thousands, the classifier will unintentionally give more importance to features having comparatively high scale values. Inorder to eliminate such a scenario and consider all features through same lens, all the

features of the dataset should be brought on the same scale. This process is called normalization and can be achieved using this setting. The software performs sklearn's MinMaxScalar normalization by transforming values of all the features in range 0 to 1.

SMOTE

The dataset in which compounds which activate a target are very less compared to compounds which do not activate it, we say there is a class imbalance in the dataset. It is important to handle class imbalance in a dataset, as without handling, even a simple classifier, which always predicts compounds as non activating, will have great accuracy performance as most of the training samples are non-activating.

To handle class imbalance, the software provides users with a setting called SMOTE (Synthetic Minority Over-sampling Technique), which is an oversampling technique that generates synthetic new samples of minority class to make the ratio of both activating and inactivating classes balanced.

Step 3: Feature Reduction

There are a total of 1875 numeric features (molecular descriptors) generated by PaDEL and other similar tools (like mordred) also generates features in thousands. The dataset of user compounds with activation behaviour generally have samples only in hundreds (like OR1A1 dataset has around only 350 samples). Theclassifier may not learn appropriately when the number of features are high compared to the number of data samples, this problem is called curse of dimensionality.

Inorder to rectify issues with high dimensionality, it is recommended to reduce the number of dimensions (features) for better predictive performance of the classifiers. It is sometimes important to filter out features which do not add much information and should only retain most important features which have discrimination power for classification. This elimination of unimportant features will not only simplify the learning of classifiers but many times increases predictive performance of the classifier.

The software provides two ways to reduce the number of features as described below.

Feature Selection

Feature selection setting allows users to retain the subset of features from the input features which are most important.

The software uses Boruta as a feature selection tool. Boruta is configured to use the default RandomForestClassifier with 100 estimators (100 trees in forest). The boruta method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilise that test.

Feature Extraction

Feature extraction setting allows users to transform high dimensional space to low dimensional space while preserving as much as information as possible.

The software uses PCA (Principal Component Analysis) as a feature extraction tool. PCA dimensionality reduction works by using less number of dimensions compared to the original dimensions to explain most of the variance in data (retain as much information as possible). Software also allows users to specify a threshold percentage (called Eigen energy) of variance that needs to be preserved on application of PCA. Recommended setting for eigen energy (threshold) is 98% (0.98), meaning the reduced number of features should still be able to explain 98% of variance in the data.

Step 4: Classification

Support Vector Machine (SVM)

Extra Tree (ET)

Logistic Regression (LR)

Gaussian Naive Bayes (GNB)

Gradient Boosting Machine (GBM)

Random Forest (RF)

Multi-Layer Perceptron (MLP)

Step 5: Validation Scheme

3-Fold Cross Validation

5-Fold Cross Validation

Leave One Out Cross Validation

Step 6: Test Set Generation

The software performs a similarity search on various databases (details of databases given below) in order to find the novel compounds which are similar to activating compounds uploaded by the user in his/her dataset. The similarity is measured using various similarity metrics (as described below). The compounds which are found similar (similarity higher than that of a threshold) are tested against the target using the trained classification models and their probability of activating the target is returned as a result. This process helps discover novel

compounds which can activate a target with high probability and are similar to activating compounds in training dataset.

The similarity comparison is performed using SMILES of chemical compounds with the help of the RDKit package. For a given chemical compound and it's SMILE, the software is first generating its fingerprint using RDKit Fingerprint. This fingerprint is then compared against the fingerprints of all compounds in user selected databases using different similarity metrics like "Tanimoto", "Dice", "Cosine" etc. at different levels of thresholds. The similarity computation is also performed with the help of the RDKit package.

Database Search

HMDB

HMDB (Human Metabolome Database (HMDB) is a database containing detailed information about small molecule metabolites found in the human body. The current version of the software consists of information of 1,14,005 metabolites downloaded as on 19th August 2019. The dataset consists of HMDB identifier (also called compound identifier), Name, SMILES of metabolites.

ChEBI

ChEBI (Chemical Entities of Biological Interest) is a database of molecular entities focused on 'small' chemical compounds. The current version of the software consists of information of 1,01,607 small compounds downloaded as on 9th March 2020. The dataset consists of ChEBI identifier (also called compound identifier), Name, SMILES of compounds.

FOODB

FooDB is the world's largest and most comprehensive database on food constituents, chemistry and biology. It consists of chemical composition data on common, unprocessed foods. The current version of the software consists of information of 26,490 food related compounds downloaded as on 21st March 2020. The dataset consists of FOODB identifier (also called compound identifier), Name, SMILES of common food related compounds.

IMPPAT

IMPPAT (Indian Medicinal Plants, Phytochemistry And Therapeutics) is the largest database on phytochemicals of Indian medicinal plants to date. The current version of the software consists of information of 1,512 phytochemicals of indian medicinal plants downloaded as on 03rd March 2020. The dataset consists of FOODB identifier (also called compound identifier), Name, SMILES of phytochemicals.

Other databases will be added to software as and when required in future updates.

Similarity Measures

Users can select one or more similarity measures to compare similarity between two

compounds. Similarity measures work by measuring similarity in molecular fingerprints of the compounds generated by RDKit.

List of all the similarity metrics along with their formulas used by software is described in the table below as an image. The fingerprints generated by RDKit are dichotomous in nature and hence the formula corresponding to dichotomous variables columns are used for similarity calculation.

Distance metric	Formula for continuous variables [a]	Formula for dichotomous variables [a]
Manhattan distance	$D_{A,B} = \sum_{j=1}^{n} \lvert x_{jA} - x_{jB} \rvert$	$D_{A,B} = a + b - 2c$
Euclidean distance	$D_{A,B} = \left[\sum_{j=1}^{n} (x_{jA} - x_{jB})^2 \right]^{1/2}$	$D_{A,B} = [a + b - 2c]^{1/2}$
Cosine coefficient	$S_{A,B} = \left[\sum_{j=1}^{n} x_{jA} x_{jB} \right] / \left[\sum_{j=1}^{n} (x_{jA})^2 \sum_{j=1}^{n} (x_{jB})^2 \right]^{1/2}$	$S_{A,B} = \frac{c}{[ab]^{1/2}}$
Dice coefficient	$S_{A,B} = \left[2 \sum_{j=1}^{n} x_{jA} x_{jB} \right] / \left[\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 \right]$	$S_{A,B} = 2c/[a + b]$
Tanimoto coefficient	$S_{A,B} = \frac{\left[\sum_{j=1}^{n} x_{jA} x_{jB} \right]}{\left[\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jA} x_{jB} \right]}$	$S_{A,B} = c/[a + b - c]$
Soergel distance[b]	$D_{A,B} = \left[\sum_{j=1}^{n} \lvert x_{jA} - x_{jB} \rvert \right] / \left[\sum_{j=1}^{n} max\,(x_{jA}, x_{jB}) \right]$	$D_{A,B} = 1 - \frac{c}{[a+b-c]}$

Here, a is the number of on bits in molecule A, b is the number of on bits in molecule B, while c is the number of bits that are on in both molecules. S denotes similarities, while D denotes distances (according to the more commonly used formula for the given metric). x_{jA} means the j-th feature of molecule A. [b]The Soergel distance is the complement of the Tanimoto coefficient. Note that distances and similarities can be converted to one another using below equation

$$similarity = \frac{1}{1 + distance}$$

View all Jobs

Paste software screenshot of “View all Jobs“ page here

Write about all job statuses
Write about how data can be viewed at different stages
Write about how logs can be viewed