**Avir.R User Instruction**

(Version 1.0, Jun 08, 2023)

Zixuan Zhang, Tao Huan\*

Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, V6T 1Z1, BC, Canada

\* Author to whom correspondence should be addressed:

Dr. Tao Huan

Tel: (+1)-604-822-4891

E-mail: thuan@chem.ubc.ca

Website: <https://huan.chem.ubc.ca/>

• Avir stands for Alignment and Integration Evaluator. It is a support vector machine (SVM)-based R program to predict the quality of metabolic peak integration in liquid chromatography-mass spectrometry-based metabolomics data.

• Avir.R script is freely available for non-commercial use.

• The instructions are given below:

**Preparation**

1. Software installation

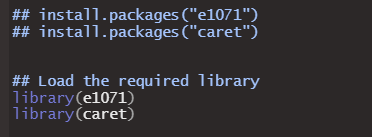
Download and install R studio following the instruction on the RStudio website (<https://www.rstudio.com/>).

1. R package installation

If the R package “e1071” and “caret” are not installed. Please run the following code then load these two packages:

install.packages("e1071")

install.packages("caret")



**Figure 1**

1. Data preparation
2. Download the R script, “Avir.R”, the SVM model, “Avir.rds”, from (https://github.com/HuanLab/AVIR.R) then save them in the folder for data processing.

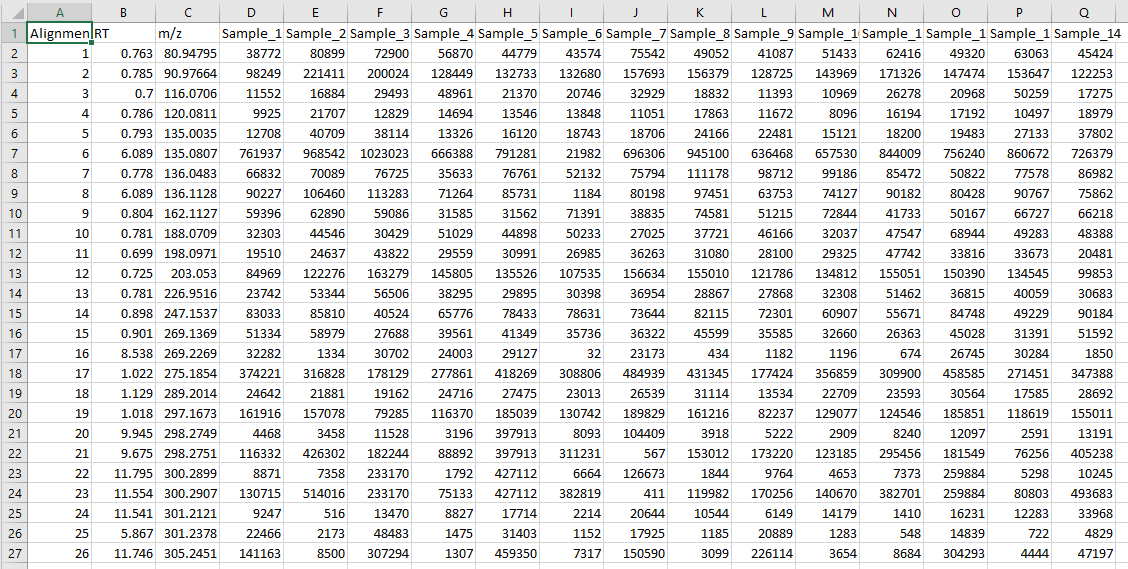
Within the same folder, prepare two .csv files, one for sample metabolite-intensities in peak area and the other for sample metabolite intensities in peak height. The content in each column should be prepared as follows (**Figure 2**).

Column 1: alignment ID

Column 2: retention time

Column 3: m/z value

Column 4 to the last column: MS signal intensities of real samples



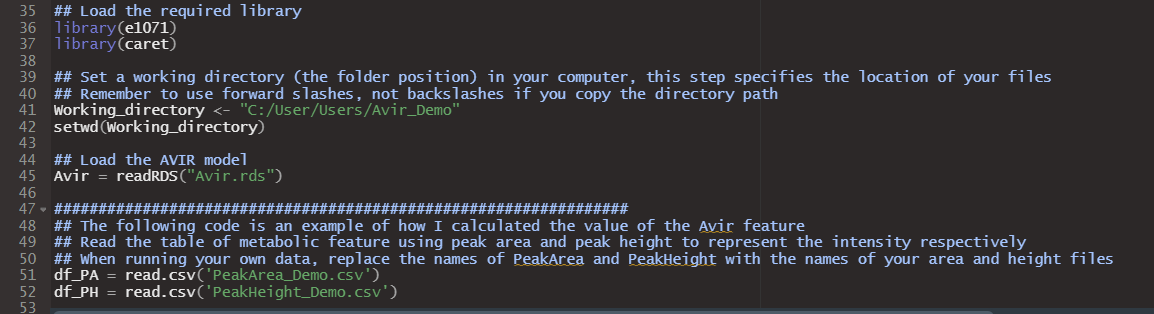
**Figure 2**

**Note:** Demo files can be found in “PeakArea\_Demo.csv” from (https://github.com/HuanLab/AVIR.R).

**Main**:

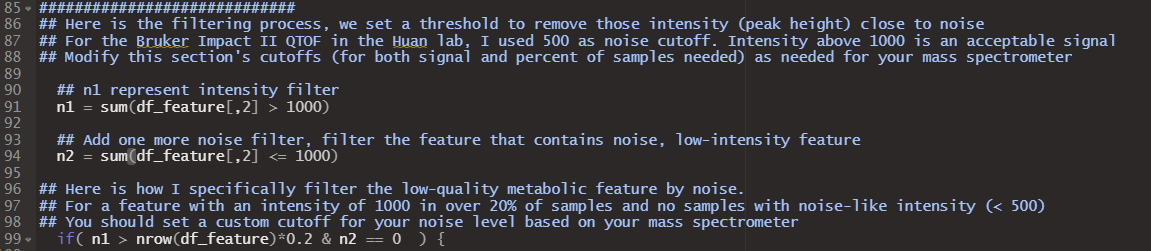
1. Set up the working directory and specify the input files (code line 41).

For demonstration, here we put the folder in the desktop named “Avir\_Demo”. Specify the name of sample metabolite-intensity in peak area and peak height (code lines 51 and 52).



**Figure 3**

1. Specify the intensity threshold and reproducibility filter for high-quality metabolic features.



**Figure 4**

To enhance the performance and prediction accuracy of the SVM model, it's essential to set appropriate intensity thresholds. These thresholds enable the exclusion of low-quality, noise-like peaks so that Avir only works on high-quality metabolic features that are more likely to be real metabolites.

Step 1: Set the intensity threshold

Lines 91 and 94 of the R script control the intensity threshold. For Bruker's Impact II QTOF mass spectrometer, we recommend setting this value to 1000 counts. Metabolic features that don't reach this threshold are considered low quality and excluded from further analysis. For the Impact II QTOF, a setting of 1000 counts is recommended. Any feature below this level will be classified as noise and subsequently discarded.

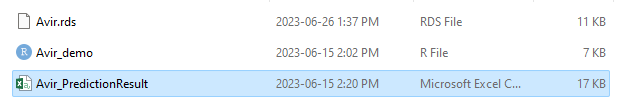
Step 2: Apply reproducibility filter

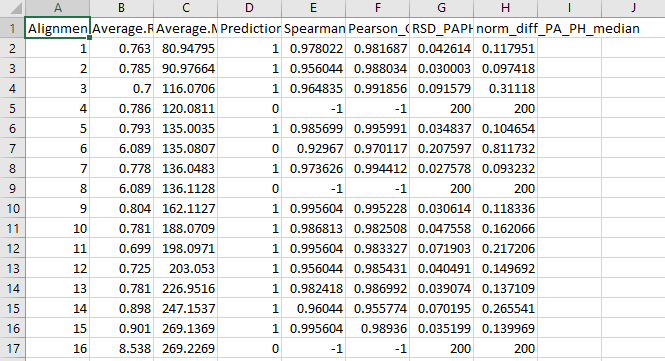
Line 99 controls the reproducibility level. To ensure the reliability of our predictions, metabolic features with low reproducibility are also filtered out. A feature will be considered high quality if at least 20% of the samples show an intensity above the set threshold of 1000 counts.

Please note that these settings are dependent on the specific experimental design, and may need adjustment depending on your research aim. The reproducibility filter can be customized by the user to fit their specific needs. By carefully adjusting these parameters, you can increase the quality of features included in the model, and potentially improve prediction accuracy.

1. Run the R script by clicking “Source” on the top right of R studio panel.
2. Check the output prediction by Avir.

The csv file named “Avir\_PredictionResult.csv” contains the prediction outcome. The first three columns of the table represent the information of Alignment ID, retention time and m/z. The four column “Prediction” represent the output of Avir. 1 means TRUE, there is no computational variation in the metabolic feature. and 0 means FALSE, there is high computational variation in the metabolic feature. The rest of four columns are the statistical properties that is used for Avir prediction.





**Figure 5**

**Note**: For slow-quality metabolic features that are not worth running machine learning prediction, extreme coefficients and values are assigned. These features can be easily recognized from the outcome as ‘-1’ is assigned to Spearman correlation and Pearson correlation (one kind of low-quality metabolic features is assigned -1, and the other kind of metabolic features that mostly contains noise is assigned 0.). In addition, RSD of PA/PH and the normalized range of PA/PH are assigned as assigned as 200. In **Figure 5**, we can see metabolic feature #4, #8 and #16 (Alignment ID, the first column) are in low quality and thus assigned extreme values.

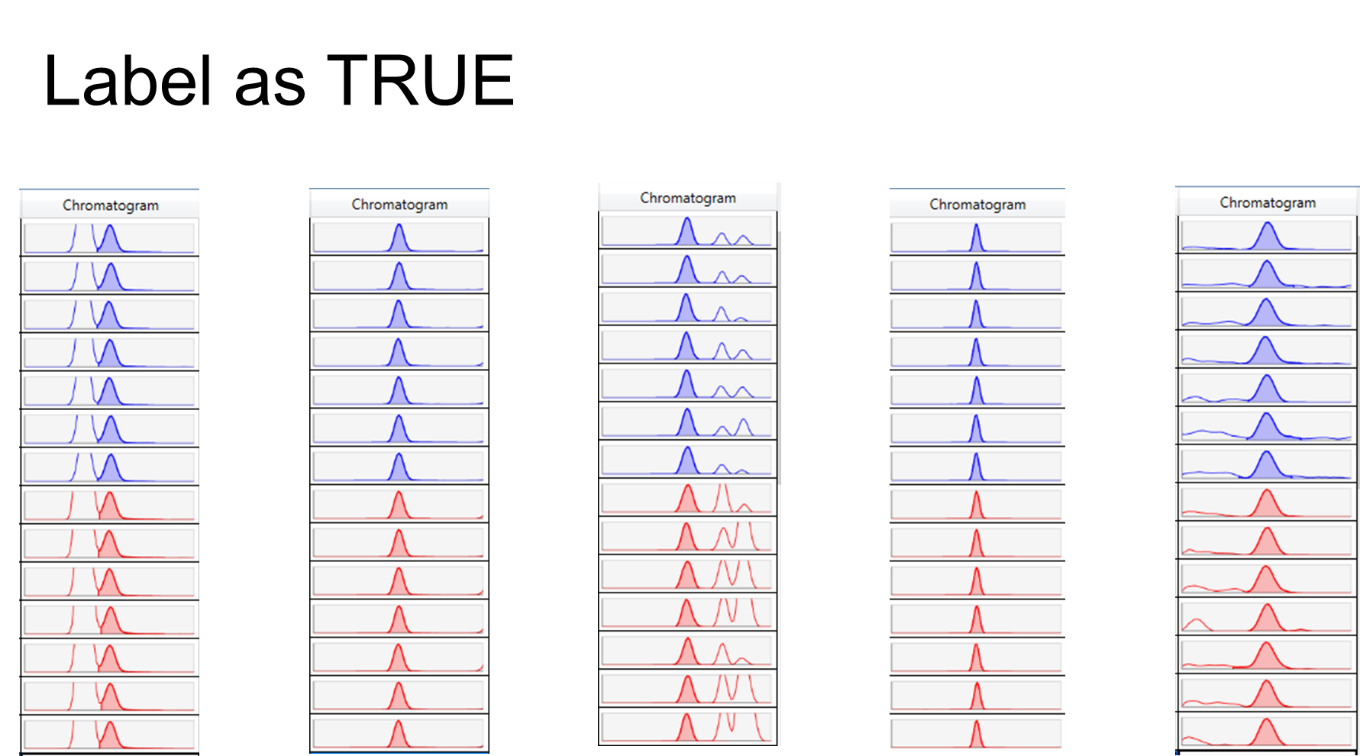
**SVM Model Development**

This section talks about how users can train their own SVM model in R for Avir applications.

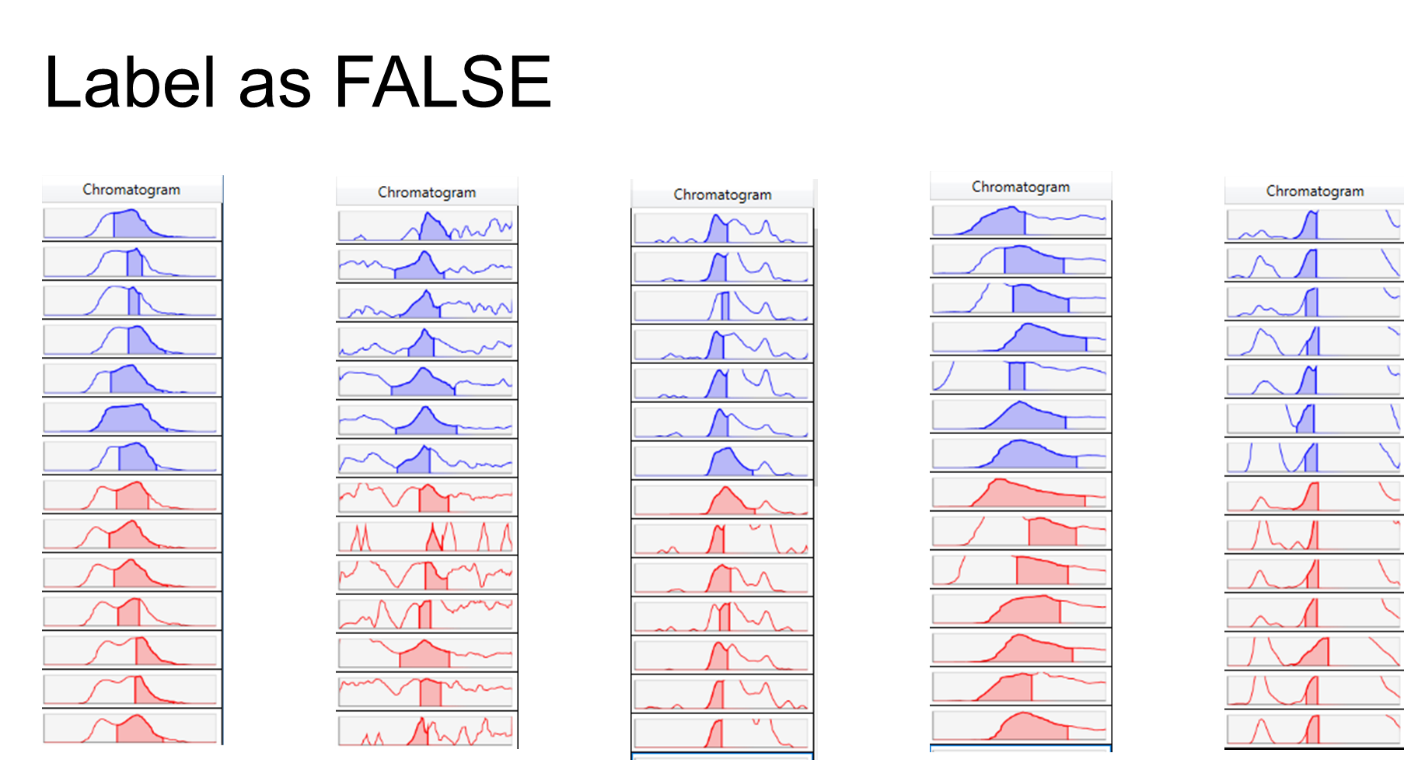
**Training data labeling**

Here we show the examples of labeling metabolic features, you can label the metabolomics data from your LC-MS platform to prepare the training dataset and test dataset.

The following figure shows five examples of metabolic features with no computational variation. These features are labeled as TRUE in the training data.



The following figure shows five examples of metabolic features with high computational variation. These features are labeled as FALSE in the training data.



**Model Training**

After collecting enough training data (it would better that size of training data is larger than 500), you can follow guidance below to generate your SVM model for prediction:

1. Download the R script, “ModelGeneration.R”, from (https://github.com/HuanLab/AVIR.R) then save them in the folder for data processing.

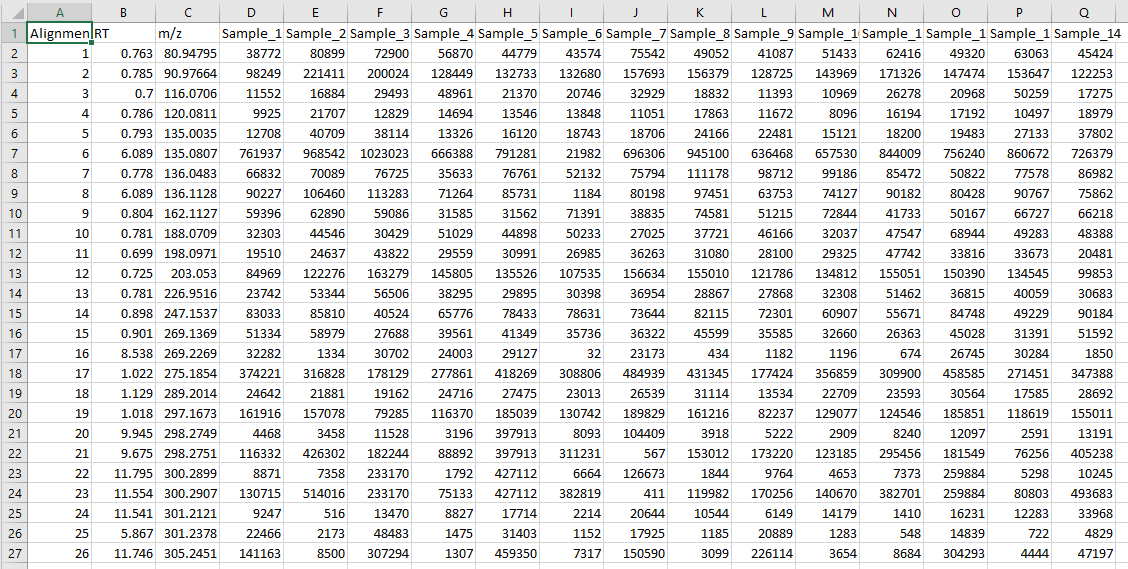
Within the same folder, prepare three .csv files, one for sample metabolite-intensities in peak area,one for sample metabolite intensities in peak height and one for the labeling results of metabolic features. The content in each column of sample metabolite-instensities should be prepared as **Figure 2**.

Column 1: alignment ID

Column 2: retention time

Column 3: m/z value

Column 4 to the last column: MS signal intensities of real samples

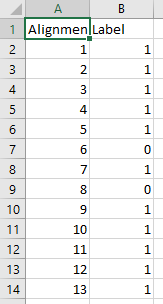


**Figure 2**

The Label.csv” file associate the True/False labels with the Alignment ID of metabolic feature. The content in each column should be prepared as follows (**Figure 6**):

Column 1: alignment ID

Column 2: label (‘1’ represents true and ‘0’ represents false)

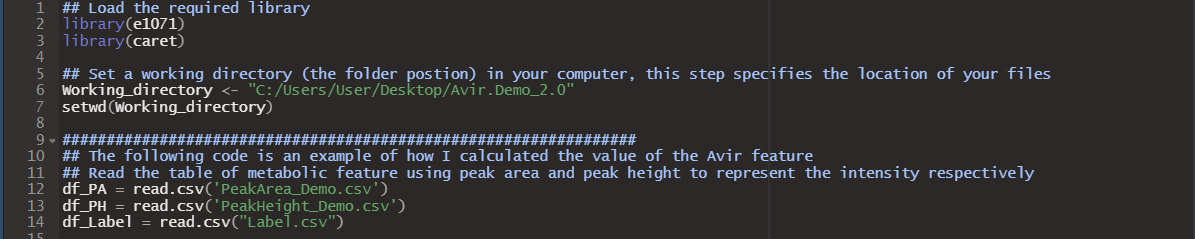


**Figure 6**

**Note:** Demo files can be found in “PeakArea\_Demo.csv”, “PeakHeight\_Demo.csv, “Label.csv” from (https://github.com/HuanLab/AVIR.R).

1. Set up the working directory and specify the input files (code line 6).

For demonstration, here we put the folder in the desktop named “Avir\_Demo\_2.0”. Specify the name of sample metabolite-intensity in peak area and peak height, and the file name of label. (code lines 12-14).



**Figure 7**

1. Run the R script by clicking “Source” on the top right of R studio panel.
2. Check the output SVM model.

Your own SVM model will be be named “SVM.rds” and saved in “rds” format in the folder. Then you can load the model for further applications.