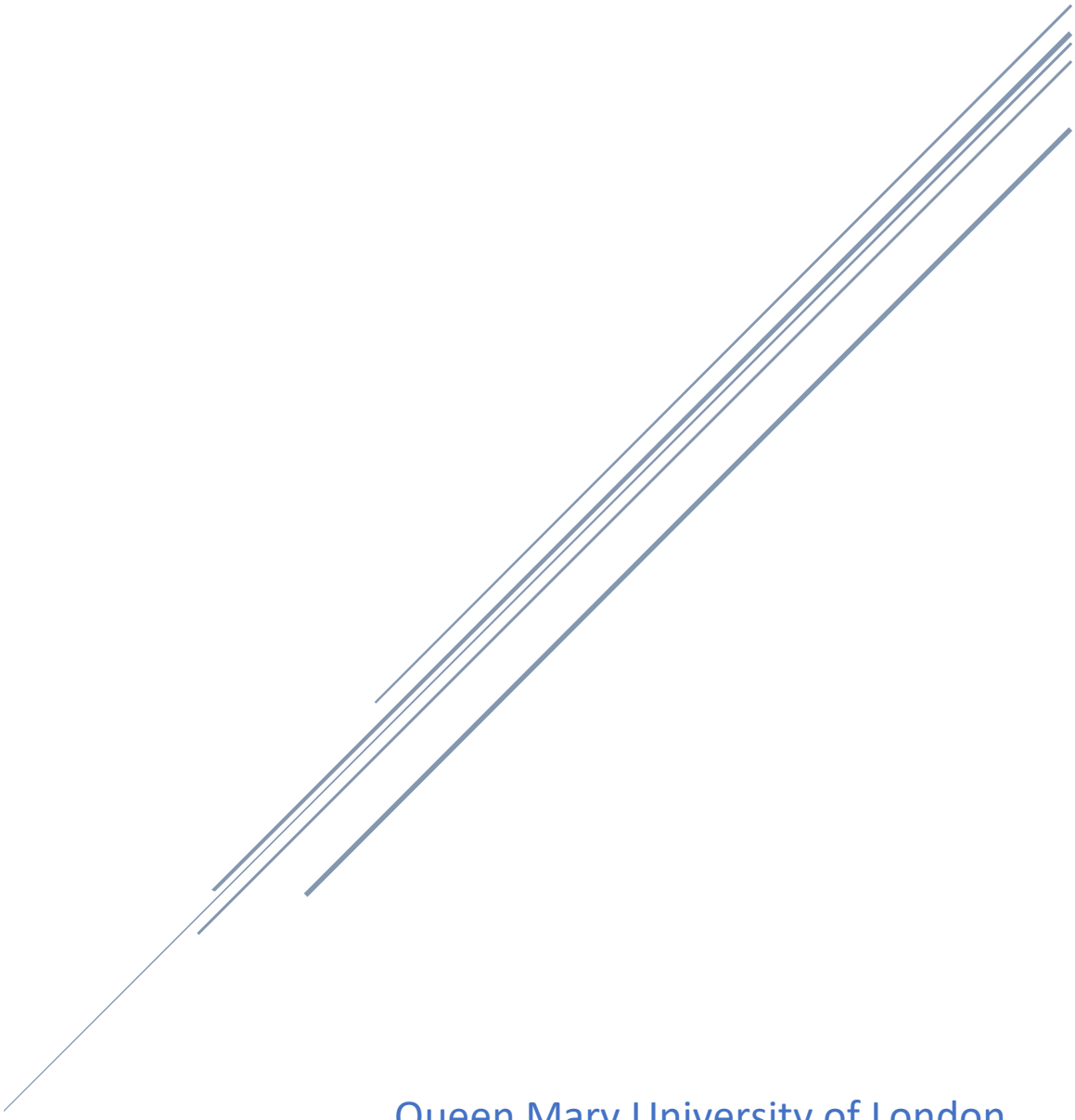
DOCUMENTATION
LIMA: LIPID MEDIATOR ANALYST

A Tool for Analysing Lipid Mediator Data



Queen Mary University of London
Zainab Haybe

List of contents

Page #

Aims

The main objective of this project is to create to create a user-friendly application for researchers to analyse lipid mediator LC-MS data and visualize results in a way to achieve a better understanding of the regulation of SPMs. This was broken down into three specific aims. The first aim is to create a pipeline for calculating lipid mediator concentrations from LC-MS data to decrease users' data processing time. The second aim is to integrate and expand previous multivariate statistics and machine learning methodologies providing users easy access to bioinformatics tools. The final aim is to build a machine learning model using Extreme Gradient Boosting for the classification of lipid mediator profiling data. This tool will also allow users to easily apply bioinformatic techniques including multivariate analysis and machine learning models to their datasets.

Using The LIMA App

Launching the Application

The application requires RStudio and the LIMA_App folder to be downloaded from GitHub (Link). The folder includes the R-script for the application including user interface and server scripts for all the pages. In addition, there is the coefficients folder with the additional information required for the application with the Cross reference coefficeints, standard curve values, pathways, fatty acid precursor, and standardized naming convention for all the mediators. The final requirement in the LIMA_App folder is the "classyfire_0.1-2.tar.gz" which is the zip file required to run the classyfire SVM package. The application can be launched by opening the LIMA_Final.R script and clicking the "Run App" button to start the application. Please ensure your folder is set up as Figure 1.

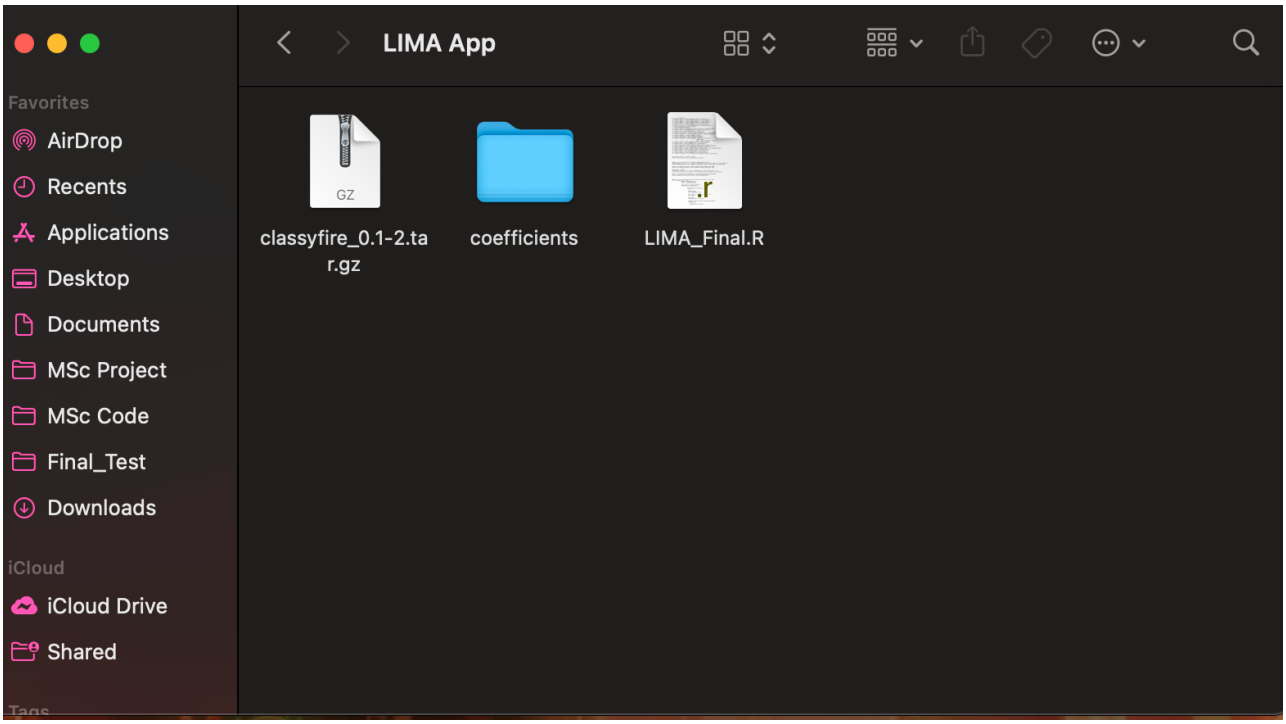


Figure 1: Contents of LIMA App folder.

Application Layout

Once the application has been launched, the first page opened is the data processing page (Figure 2A). At the top of the application is a navigation bar which has the name of the application "LIMA" (Lipid Mediator Analyst) and the tabs for each of the pages including Data Preparation, Multivariate Statistics, and ML Models (Figure 2A). The Data Preparation tab has drop down for data processing and Data Formatting pages (Figure 2B). The Multivariate Statistics has a drop-down navigation with the PCA/PLSDA, Differential Analysis, and Correlation Analysis pages for each of the three multivariate statistical tests (Figure 2D). The ML Models has a drop-down menu for Machine Learning, Optimize ML Model, Build ML Model, and Run ML models where machine learning models can be built and optimized on a training dataset and run on a validation dataset (Figure 2D). Each of the pages has a similar layout with side bar on the left side panel and a main panel on right side of each page. The side panel is where input for the page is, with the first being a file input where the user can upload a file with the mediator concentrations for each of the sample and additional information that is specific to each page. The second and third input are the checkbox input for if the file contains a

header and the separator format of the file respectively. The remaining inputs are specific for each of the pages based on their function. The main panel is where the outputs for each page are shown which include the plots, tables, and the download buttons to obtain the data. The main panel contains multiple tabs to separate the different outputs with there being one to two outputs on each tab.

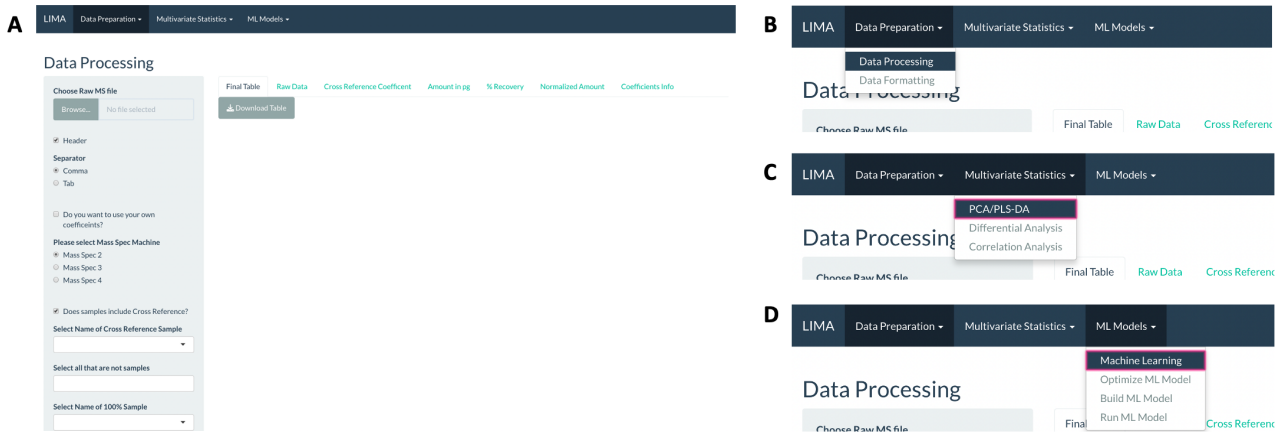


Figure 2: A) Layout of SPM Analyst with the first page once the application is launched from RStudio. B-D) The navigation bar and each page for the drop down is shown.

Data Preparation

Data Processing

The Data Preparation tab on the navigation bar directs users to the Data Processing page. The sidebar on the right side is the user inputs for the data processing page (Figure 3). The file input for the data processing page is the table from the Mass Spectrometry data analysis software with the area for each mediator. This is followed by a checkbox for the user to select if they want to use their own coefficients. The default is that the checkbox is not checked (FALSE) which leads to the buttons where the user can select which mass spectrometry machine was used to obtain the default coefficients (Figure 3A). If the check box is selected, the input box changes to enable users to upload their Cross Reference and Standard Curve coefficients file with each file

having a header and separator input (Figure 3B). In addition, the user can download example Cross Reference and Standard Curvet files to ensure the files is in the correct format. The fourth input is a checkbox if there is a cross reference sample included, with it checked in the default setting (TRUE). When it is clicked, there is a box to select the name of the cross-reference sample (Figure 3C). If it is not selected, the input box for selecting the cross refence name does not appear (Figure 3D). This is followed a box used for selecting all those in the sample columns that are not samples to be studied. Then there is in a box for selecting the sample that is the 100%, which has not gone through the extraction process. The next input it to select the Area to filter to zero which has a default value of zero, and the user can select the threshold of their choice. Then there are two checkboxes where the user can select if the samples are all the same volume/weight or different volumes/weights. If samples are all the same volume/weight, an additional numeric input box will appear for the user to input the sample volume/weight and the value to standardize the concentrations (Figure 3C). If the samples are different volumes/weights, there is a sample list file that the user can download to add the specific volume/weight for each sample and the user can reupload that file for analysis (Figure 3D). Once all the inputs have been selected, the user clicks the action button "Process" for server functions to convert the area into concentrations.

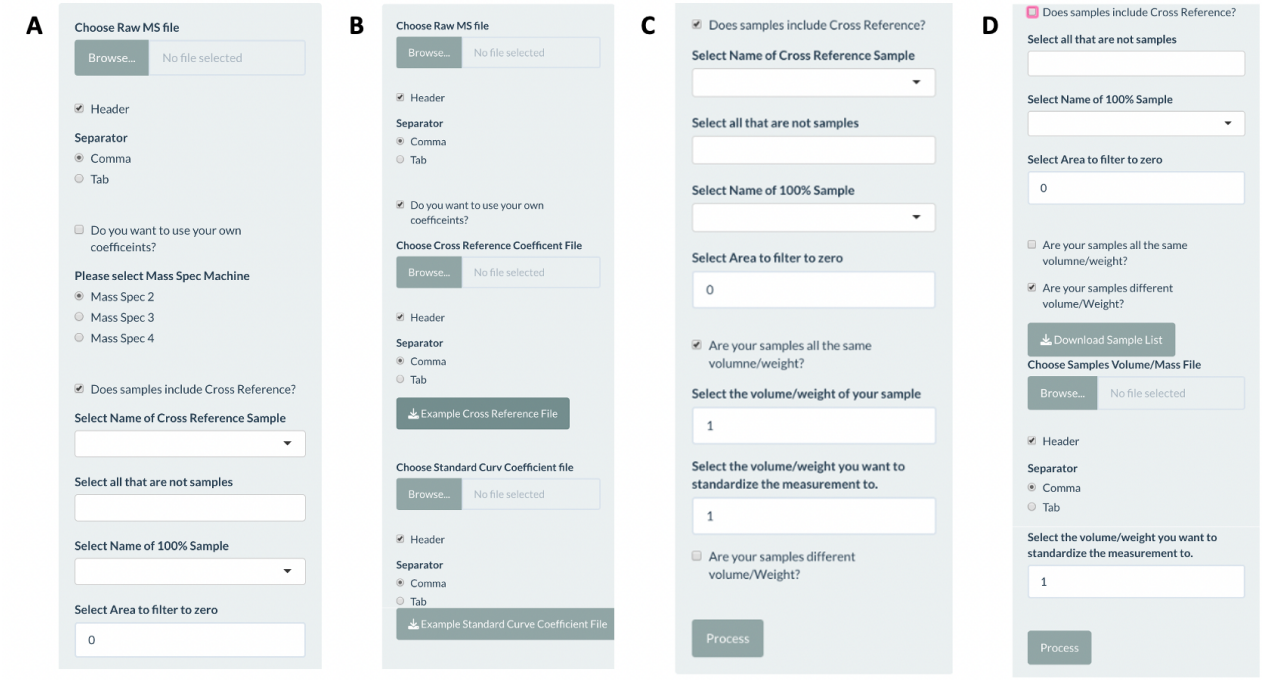


Figure 3: Sidebar panel for the Data Processing page under the Data Preparation tab. A) The default inputs using the default coefficients for the calculations. B) The inputs if the user wants to use the own coefficients. C) The inputs if the user wants to include a cross reference samples and the volume/weight of samples are the same. D) The inputs if the user does not want to include a cross reference samples and the volume/weight of samples are different.

The output is separated into multiple tabs with the first tab being the final concentration table which has the mediators as the columns, samples as the rows, and the concentration as the values (Figure 4A). The next tab contains the raw data file with the mediators as columns, samples as rows, and the raw area as the values (Figure 4B). The third tab is a table with the area after the cross-reference correction factor has been applied (Figure 4C). The fourth tab is a table with the amount in picograms of the metabolites (Figure 4D) and the fifth tab is the precent recovery of the internal standards for each of the samples (Figure 4E).The sixth tab has a table with the normalized amount after the sample loss during the preparation process has been considered (Figure 4F). The first six tabs all have download buttons where the user can download the tables as a csv and the table's displays are interactive where the user can scroll across the columns and rows. The final tab has the coefficient information with the coefficient values that were used in the calculations (Figure 4G). If there is error that occurs, such as the wrong file was uploaded, an error messages appears instead of the tables with suggestions on how to resolve the error (Figure 4H).

Figure 4: Main panel for the Data Processing page under the Data Preparation tab once inputs have been completed. The outputs are separated into 7 tabs called A) Final Table, B) Raw Data, C) Cross Reference Coefficient, D) Amount (pg), E) % Recovery, F) Normalized Amount, and G) Coefficients Info. H) The error message when the inputs are incorrect and an error occurs.

Data Formatting

The second page under the data preparation tab is the Data Format tab. The first input is the final concentration file from the Data Processing page (Figure 5A). Then there is a checkbox asking if users want to select which mediators they want to include (Figure 5A). A drop-down list of all the mediators with the transition code is included and the user can click to select which mediators they want to process (Figure 5A). The outputs are separated into two tabs, the first being the Statistics Table, where mediators are renamed in the correct format without the transition codes (Figure 5B). The second tab output is the Machine Learning Tab where a table similar table as the statistics page is outputted with the addition of the first row that contains the fatty acid precursor of each mediator (Figure 5C). Both tables can be downloaded and used as the input for the multivariate statistics and the machine learning pages respectively.

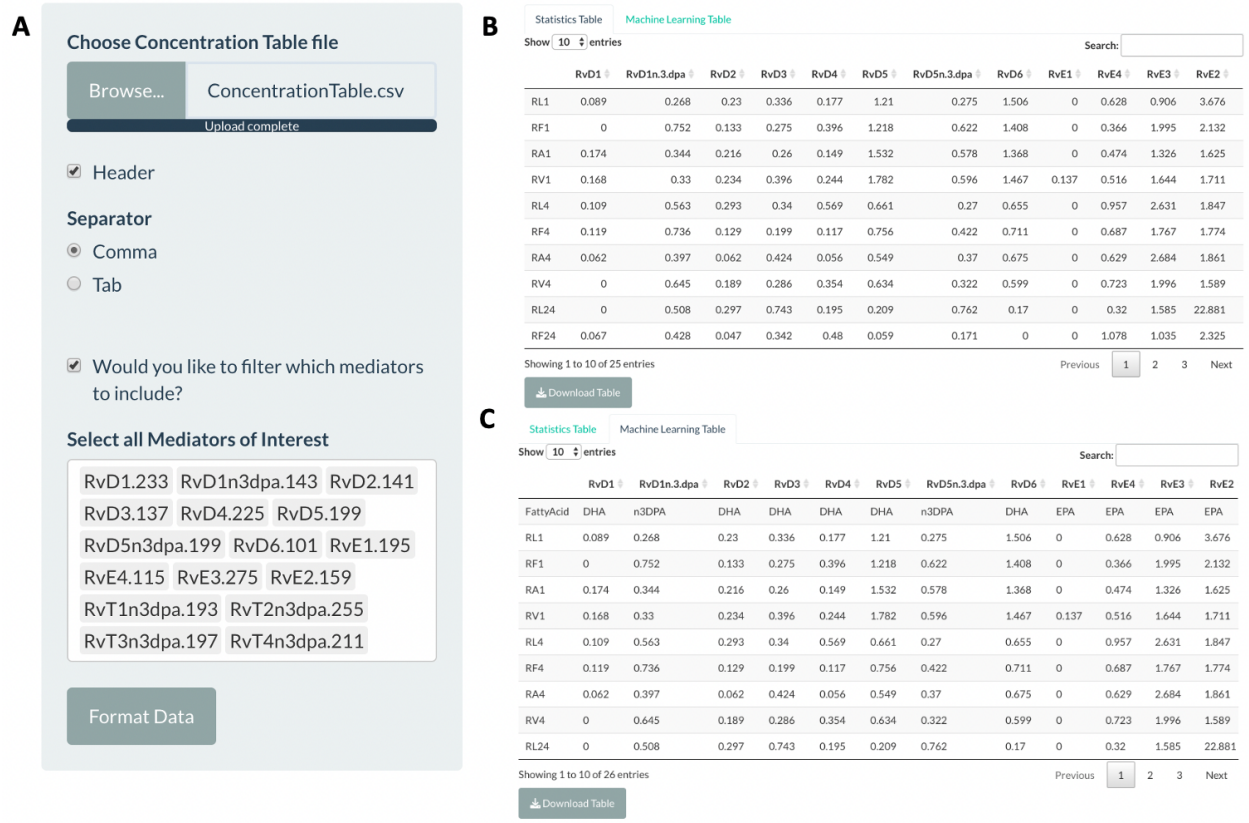


Figure 5: A) The sidebar panel for the data formatting page with the inputs. B) The outputs for the first tab of the main panel with the statistics table and download button. C) The outputs for the second tab with the machine learning table and download button.

Multivariate Statistics

PCA/PLSDA

The Multivariate Statistics tab on the navigation bar has a drop-down menu, with the first of three pages being the PCA/PLSDA page. The file that is uploaded needs to contain the mediator concentrations for each of the samples as well as the group information. The next input after the file input is the sample location, which specifies if the samples are the row names or the column names. The following inputs are the sample names and group information column names, where the user selects the name of the column that contains the sample and group information. The names are automatically generated once the file has been uploaded onto the page. The final input is to select either the PCA or PLSDA test, followed by the calculate button to run the test. (Figure 6A)

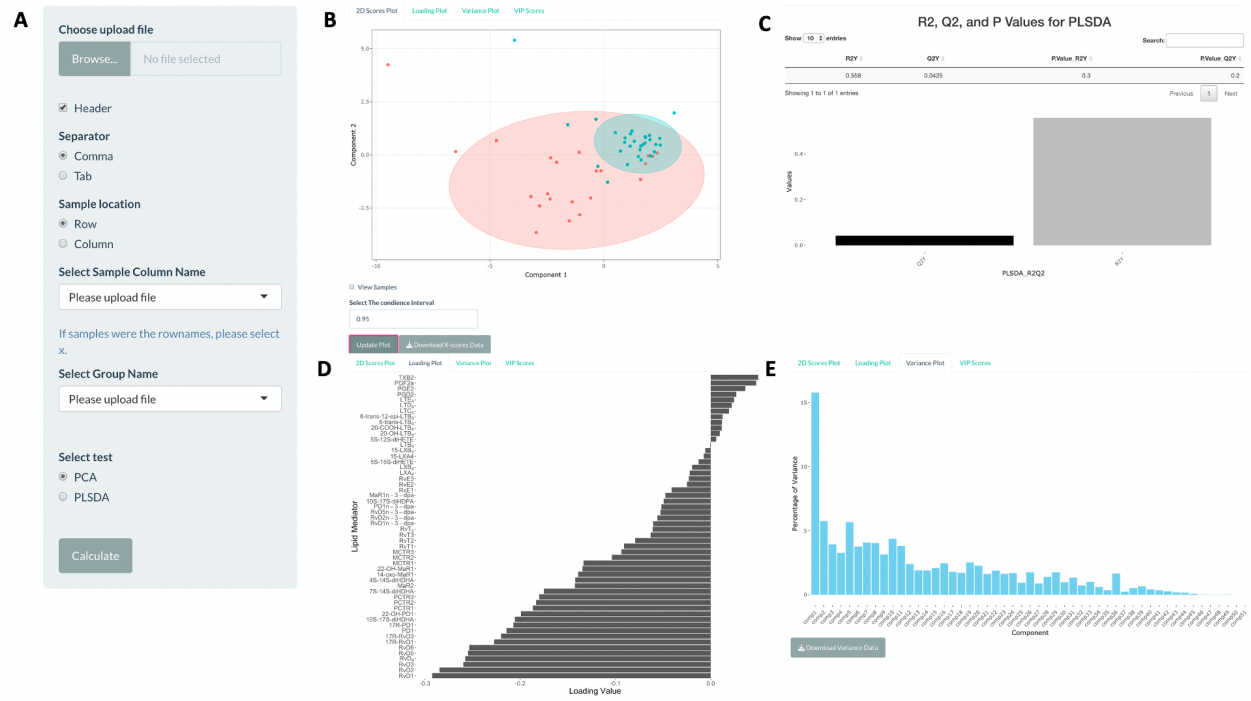


Figure 6: The sidebar panel and main panel for the PCA/PLSDA page under the multivariate statistics drop down. A) The sidebar panel with the inputs. B) The first tab of the main-panel, 2D-Scores Plot with the PCA/PLSDA scores plot of component 1 vs component 2. C) The R2/Q2 and p-values table for PLSDA model as well as the R2Q2 plot. D) The second tab with the loading plots for the mediators. E) The third tab with the percentage of variance explained for each component.

The PCA/PLSDA page has 4 tabs as the outputs on the main panel on the right of the page. The first tab is the 2D Scores Plot, with an interactive plot of the scores for the first two components for each of the samples (Figure 6B). The sample names can be shown or removed using the sample names checkbox and the confidence interval for the ellipses can be adjusted using the numeric input. If the PLSDA test is run, the R2, Q2, and P-values table is shown with the R2Q2 bar chart below the scores plot (Figure 6C). The second tab is the Loading Plot, where the weights for each of the mediators in differentiating between the groups is displayed as a bar graph (Figure 6D). The plot was created using ggplot2 which is not interactive, so there is a download button for the user to download the plot. The third tab, Variance Plot, displays an interactive bar plot with the percentage of variance for each of the PCA/PLSDA components (Figure 6E). The interactive plot can be clicked to see the information of each bar, zoomed in, and download.

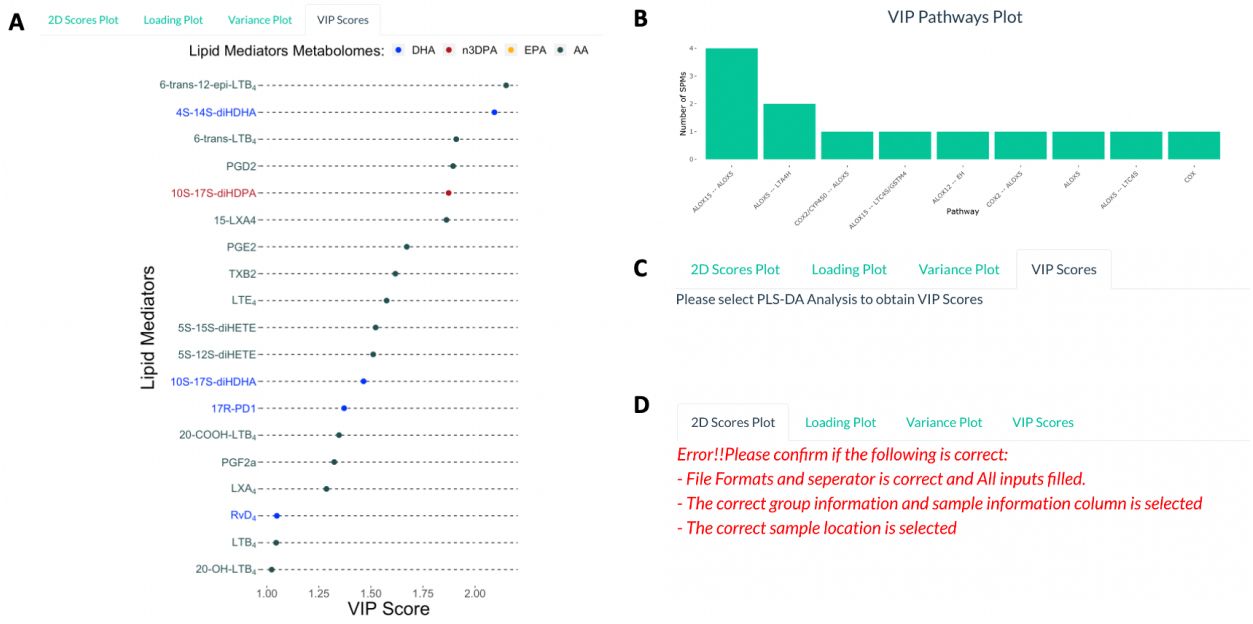


Figure 7: The final tab on the main-panel for the PCA/PLSDA page under the multivariate statistics drop down. A) The Variable Importance in Projection (VIP) plot including mediators with the VIP score greater than 1. B) The VIP pathways plot showing the pathways represented in the mediators with VIP score greater than 1. C) The message that appears when PCA is run. D) The error message when an error occurs.

The final tab is the Variable Importance in Projection (VIP) Scores tab, containing a plot with the variance importance scores for the mediators used to build the PLSDA model (Figure 7A). The plot is colour coded for each of the fatty acid precursors and the plot only shows those

values with a mean decrease score greater than one. The plot is not interactive and can be downloaded using the download button and the bottom of the page. Below the VIP Scores plot is the pathways plot, which is a bar chart of the number of SPMs for each pathway with a VIP Score greater than one (Figure 7B). If the PCA test is run, a VIP plot and pathways chart cannot be generated, and a message appears informing the users to run PLSDA to obtain the VIP plot (Figure 7C). Each tab has a download button where the user can download the data used to create the plot including the scores for each sample, the weights for the mediators, the percentage of variance of each component, and the VIP scores for the mediators. In addition, all tabs include error messages when the inputs are incorrect, or an error occurred during the analysis process (Figure 7D).

Differential Analysis

The second page in the Multivariate Statistics drop down is the Differential Analysis, which contains the t-test and the Mann-Whitney U test. The file requires the mediator concentration for each of the samples and the group information which is broken down into two groups. Following the file upload is the sample location input, to select if the samples are on the rows or the columns. Next the multivariate Normality test is selected which can either be Mardia, Henze-Zirkler, or Royston tests. The final input is the selection of Group A and Group B to be compared which is automatically populated from the Group information column. The action button at the bottom allows the calculation to be run (Figure 8A).

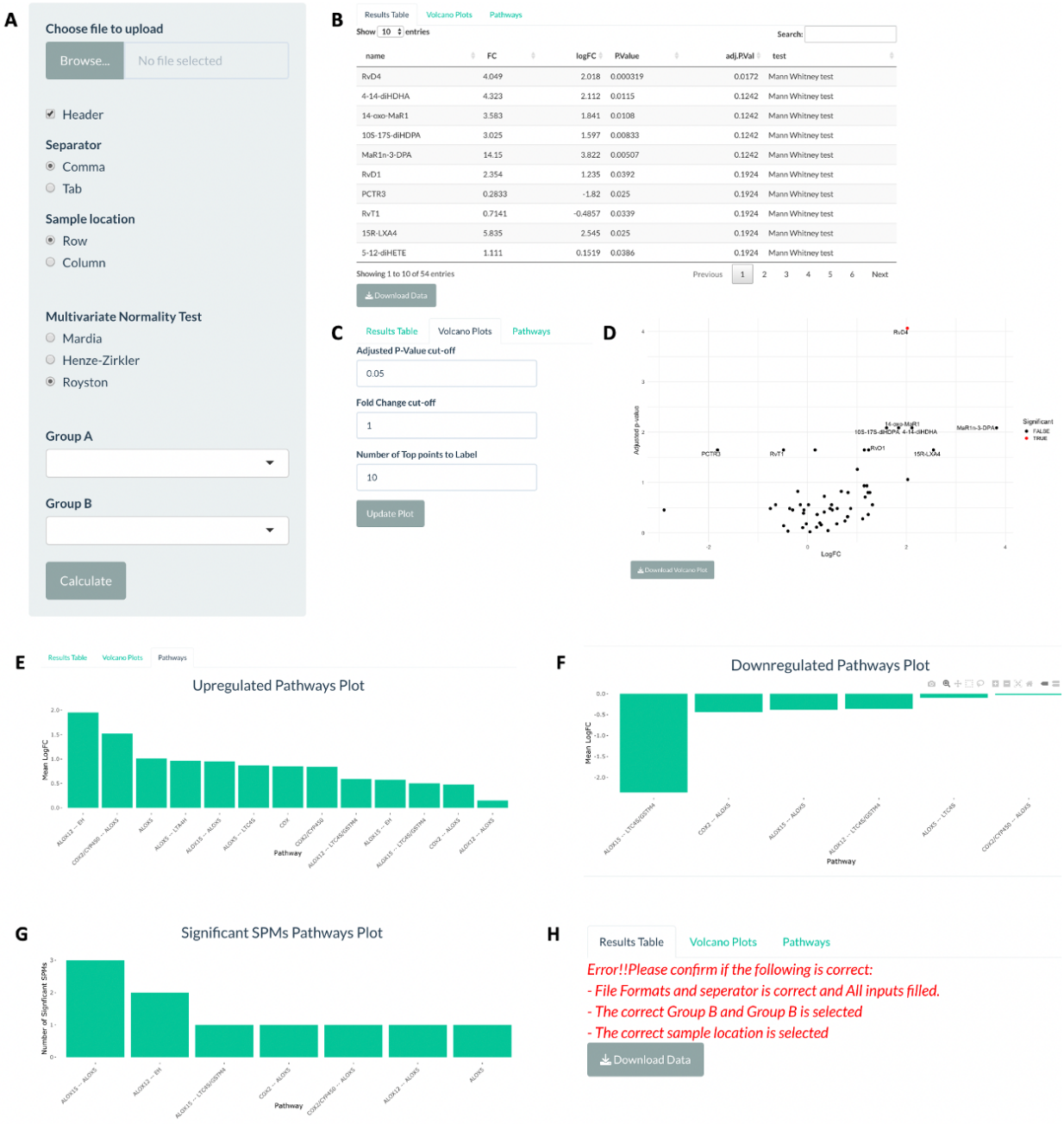


Figure 8: The sidebar panel and main panel for the Differential Analysis page under the multivariate statistics drop down. A) The sidebar panel with the inputs. B) The first tab of the main-panel, Results Table with the mediators, fold change, p-values and test used. C) The second tab with the filters for the volcano plot including p-value and fold change significance cut-off. D) The volcano plots. E) The third tab with the upregulated pathways bar chart. F) The downregulated pathways bar plot. G) The bar plot with the number of mediators with p-value less than 0.05 for each pathway. H) The error message when the analysis does not work.

The main panel for the Differential Analysis page contains three tabs, Results Table, Volcano Plot, and Pathways. The Results Table tab contains a table with the name of the mediator, Fold Change (FC) value, logFC, p-value, adjusted p-value, and the t-test that was used

which can either be the Mann-Whitney U test or the T-test (Figure 8B). There is a download button below the table where the user can download the table as a csv file. The Volcano plot tab has the volcano plot with the logFC on the x axis and the adjusted p-value on the y axis. The black indicates points that do not meet the significant threshold and the red points are the significant threshold (Figure 8D). Above the plot, are three inputs where the user can change the adjusted p-value cut off and the fold change cut-off which changes the thresholds of significance (Figure 8C). The third input is the Number of top points to label, where the numeric input indicates how many points to label. The default value is 10 but it can be adjusted from 0 to all the points. Between the three inputs and the plot is the Update Plot button used to display the plot with the changed inputs. Below the plot is a download button where the volcano plot can be downloaded. The Pathways tab contains three bar charts, the first being the mean log fold change for each of the upregulated pathways (Figure 8E). The second bar chart is the mean log fold change for all the downregulated pathways (Figure 8F). The final bar chart is the number of mediators for each pathway that has a significant p-value of less than 0.05 (Figure 8G). The three bar charts are interactive and can be downloaded as a png file. Once an error occurs in the analysis, an error message appears with suggestions on how to correct common errors (Figure 8F).

Correlation Analysis

The final page under the Multivariate Statistics tab is the Correlation Analysis page where spearman's correlation is run on the uploaded dataset. The first input on the side panel is check box to select if the data is all in one table. If the check box is clicked, then the side panel has one file upload input and selecting if the sample location if it is on the row and column (Figure 9A). In addition, there are six select input boxes that is automatically populated with the column names to select the columns that contain the following information: sample names, condition, Group A start, Group A end, Group B start, and Group B end (Figure 9A). Then there is an action button where the user can process the data and obtain the results. If the first checkbox is not clicked, the side panel changes with two file inputs, the first for Group A's file and the second for Group B's File (Figure 9B). For each of the files, the user selects if the sample names are located on the rows or columns and the condition information column name.

There is a “Calculate” button at the bottom for the spearman’s correlation test to be run and the results to be displayed.

Figure 9) The sidebar panel and main panel for the Correlation Analysis page under the multivariate statistics drop down. A) The sidebar panel with the inputs when the data is all in one file. B) The inputs when the data is separated into two files. C) The first tab of the main panel with the correlation table. D) The second tab of the main panel with the threshold filters. E) The correlation plot on the second tab of the main panel. F) The error message for the correlation analysis page.

The main panel of the Correlation Analysis page has two tabs, the Correlation Table and Correlation plots tab. For the correlation table tab, there is an input where the user can select which of the conditions, they want to show the correlation table. Once the condition has been selected and the view table button has been clicked, the table will be shown with Group A as the rows, Group B as the columns, and the Spearman's correlation as the values (Figure 9C). The user is also able to download the table using the download data button. The correlation plots tab has two inputs, the first to select which condition to plot and the second is the p-value cut-off (Figure 9D). The correlation plot will be displayed once the View plot button has been clicked and the download plot button is used to download the plot as a png file (Figure 9E). In the case of an error, the error message appears in the main panel with trouble shooting advice (Figure 9F).

Machine Learning

Machine Learning Model Page

The first drop down page from the ML Models menu is the Machine Learning models page. The first input in the side bar is the file upload where the file has the mediators as columns, the samples as rows, and the first row being the fatty acid family (Figure 10A). The second input is the separator of the file format followed by a select input box to select the column with the group information (Figure 10A). There is also a download button to download an example machine learning file. The final input is a checkbox with each of the 5 machine learning modes including Random Forests, Support Vector Machine, Bayesian Classifier, Elastic Net Regression, and Extreme Gradient Boosting (Figure 10A).

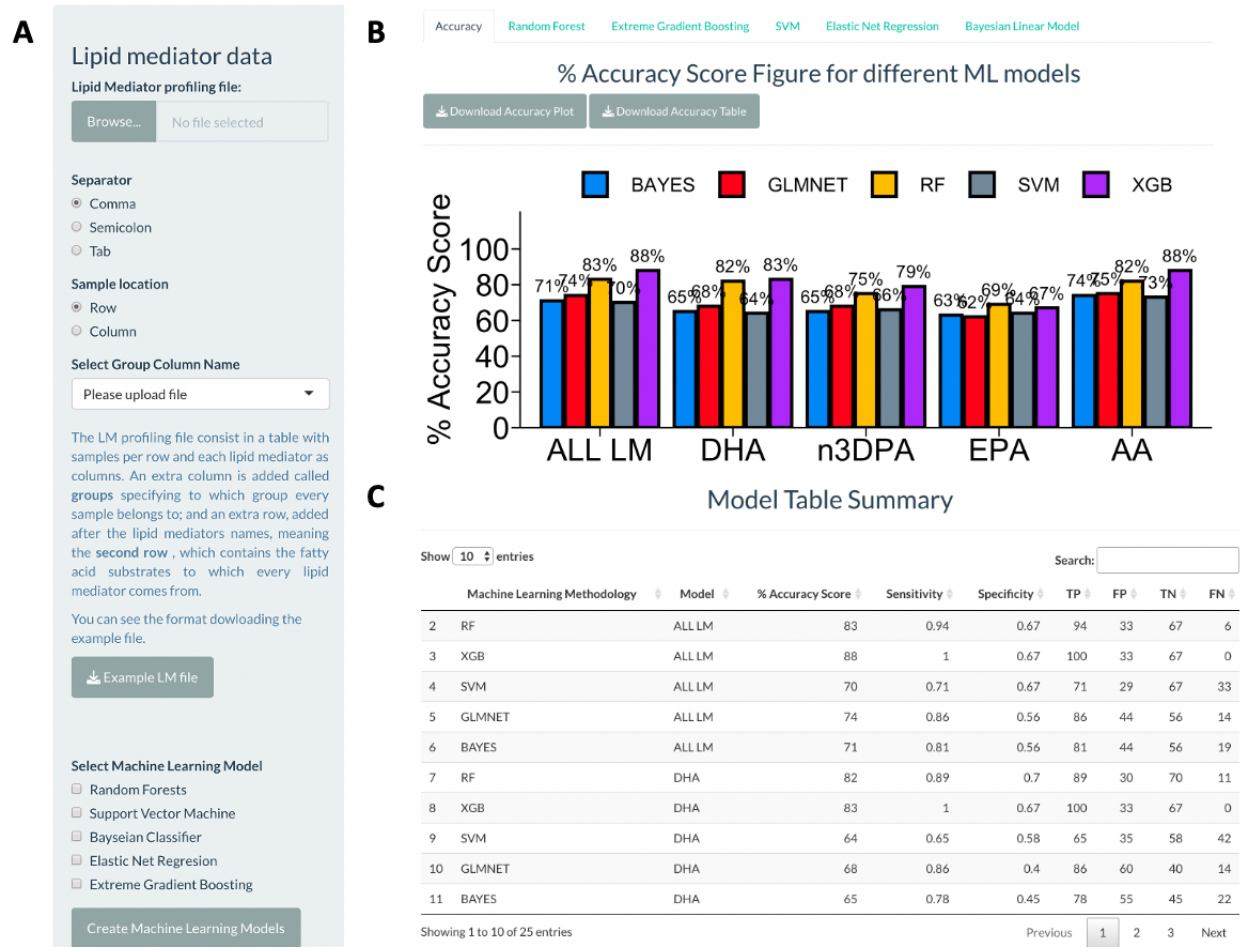


Figure 10: The sidebar panel and main panel for the Machine Learning page under the ML Models drop down. A) The sidebar panel with the inputs. B) The Accuracy tab's first output of the % Accuracy plot for all the different Machine learning models. C) The Accuracy tabs second output of the model summary table.

The first tab in the main panel is the Accuracy tab displaying a bar graph with the models on the x-axis and the average accuracy on the y-axis (Figure 10B). The models which were selected to run will be shown, with every machine learning method being repeated for all the lipid mediators and each of the fatty acid precursors. There is also a Model Summary table below the accuracy plot containing the machine learning method, model name, accuracy, sensitivity, specificity, true positives, false positives, true negatives, and false negatives for each model (Figure 10C). There are two download buttons at the top of the page, one for downloading the accuracy plot and the second for the model summary table. The remaining four tabs has specific plots for the random forest, extreme gradient boosting, support vector machine, and elastic net regression methods.

The Random Forest tab contains two sub-tabs called Parameters and Importance. The parameters tab contains a line graph with the number of trees vs the average percent accuracy (Figure 11A). There is a plot for each of the models, depending on the types of mediators included. The Importance tab contains the importance plots with the mean decrease accuracy vs the lipid mediators (Figure 11B). For each model there is one plot, and all the mediators are colour coded based on the fatty acid precursor. The plots and models can be downloaded individually, and the user can select which model or plots they would like to download (Figure 11C).

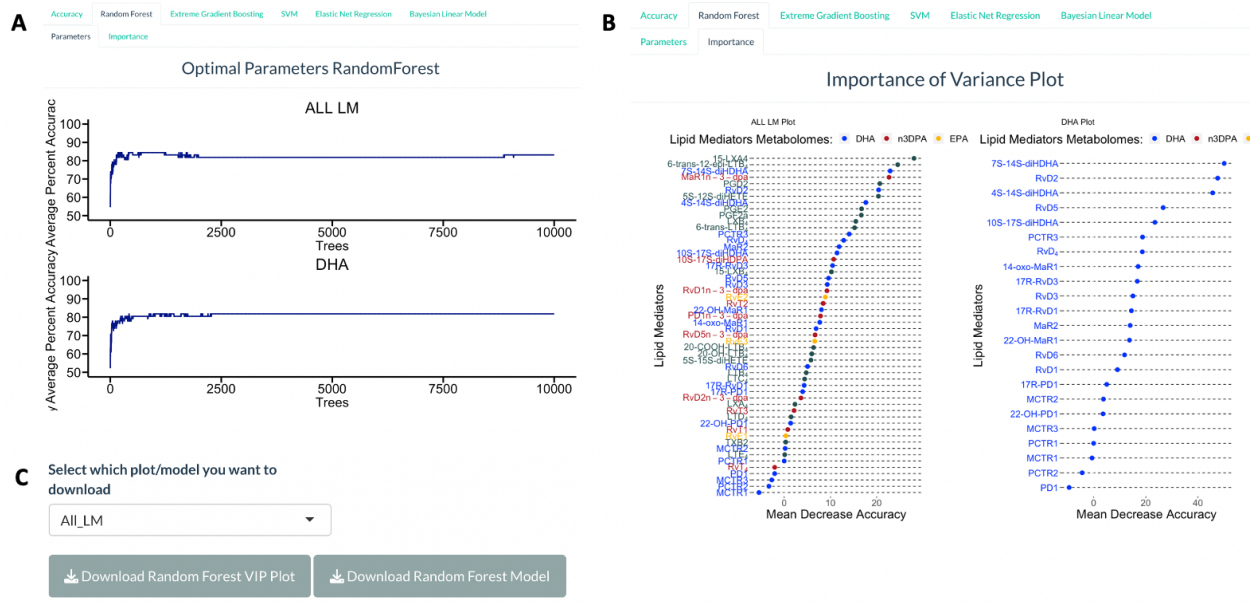


Figure 11: The random forests tab on the main panel of the machine learning models page. A) The parameters sub-tab of the random forest tab with optimal parameters plots for the random forest models. B) The second sub tab with the features of importance plots for random forests. C) The download buttons for the importance plot and random forest models.

The next tab is the Extreme Gradient Boosting tab which also contain the two Parameters and Importance sub-tabs. The parameters tab for the extreme gradient boosting models contains a table with the model information including the model's name, accuracy, specificity, sensitivity, true positive, false positive, true negative, false negative, and hyperparameters (Figure 12A). The hyperparameters include the number of rounds, learning coefficient (eta), maximum tree depth, gamma, column sample rate, subsample ratio, and minimum child weight. In addition to the table, the parameters tab contains a line plot with the number of iterations vs the test error rate for the top 5 models with the highest accuracy (Figure 12B). The plot is interactive using **plotly**,

so the user can select which of the models they would like to see and download the plot. The importance tab for the extreme gradient boosting method contains the importance plots with the gain for each of the metabolites shown for each of the models (Figure 12C). The plots and models can be downloaded individually, and the user will select which model they would like to download (Figure 12B).

Figure 12: The Extreme Gradient Boosting tab on the main panel of the machine learning models page. A) The parameters sub-tab of extreme gradient boosting tab with top 5 models and the parameters used to build the models' table. B) The optimal parameters plot for the extreme gradient boosting models and the download button. C) The second sub tab with the features of importance plots for the extreme gradient boosting models.

The fourth tab is the SVM tab which contains the optimal parameters plots for the support vector machine model. The plot is a scatter plot with number of ensembles vs the average percent accuracy (Figure 13A). The plots can be downloaded using the download button which downloads all the plots into one file. The fifth tab is the Elastic Net Regression tab which contains the optimal parameter plot for the elastic net regression models. On the x-axis is the alpha, the y-axis is the average percent accuracy, and the three lines indicate the different lambda values (Figure 13B). The plots can be download using the download button which saves the plots as a png file. The final tab is the Bayes GLM tab which does not contain any plots, but the user can download the model (Figure 13C). The models for SVM, Elastic Net Regression and Bayes

GLM can be downloaded individually, and the user will select which model they would like to download.

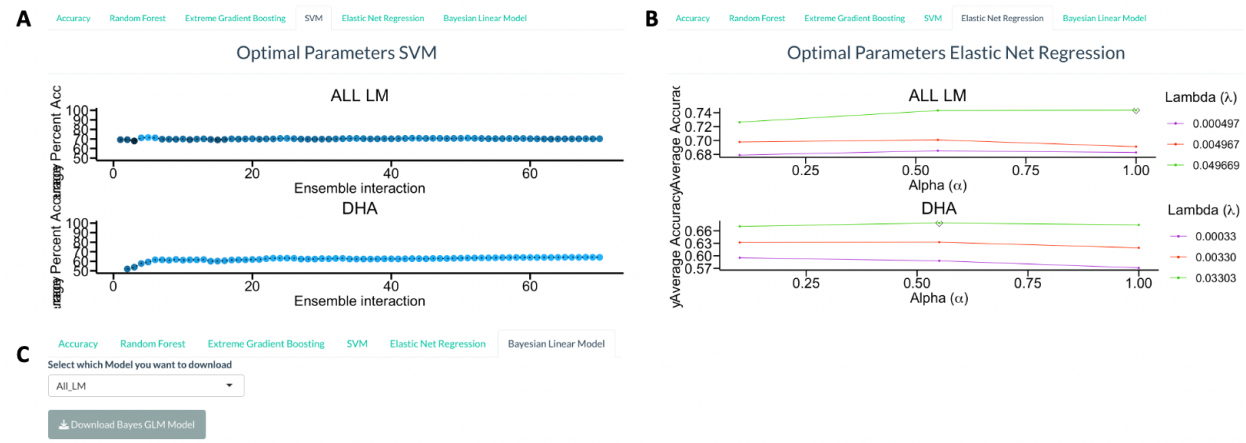


Figure 13: The SVM, Elastic Net Regression and Bayesian Linear Model tabs on the main panel of the machine learning models page. A) The SVM tab with the optimal parameters plots for the SVM models. B) The Elastic Net Regression tab with the optimal parameters for the elastic net regression models C) The Bayesian Linear Model tab with the download button for the Bayesian model.

Optimize Model Page

The second page under the ML Models drop down is the Optimize ML Model page where the user can change the parameters of the machine learning model. After the file input, there are checkboxes for each of the 5 machine learning models and once selected, specific parameters that can be changed are shown for that model. The first checkbox is for random forests and once selected, a numeric input appears where the user can select the number of trees (Figure 14A). The following checkbox is for extreme gradient boosting, and once selected an input prompting for the number of rounds appears (Figure 14A). In addition, there is an input to select which of the top 5 models the user wants to build for each of the fatty acid families. The third model is support vector machine, and once selected the user can input the number of ensembles and the number of bootstraps (Figure 14B). The final two models, elastic net regression and Bayesian analysis requires users to select the number of ensembles (Figure 14B).

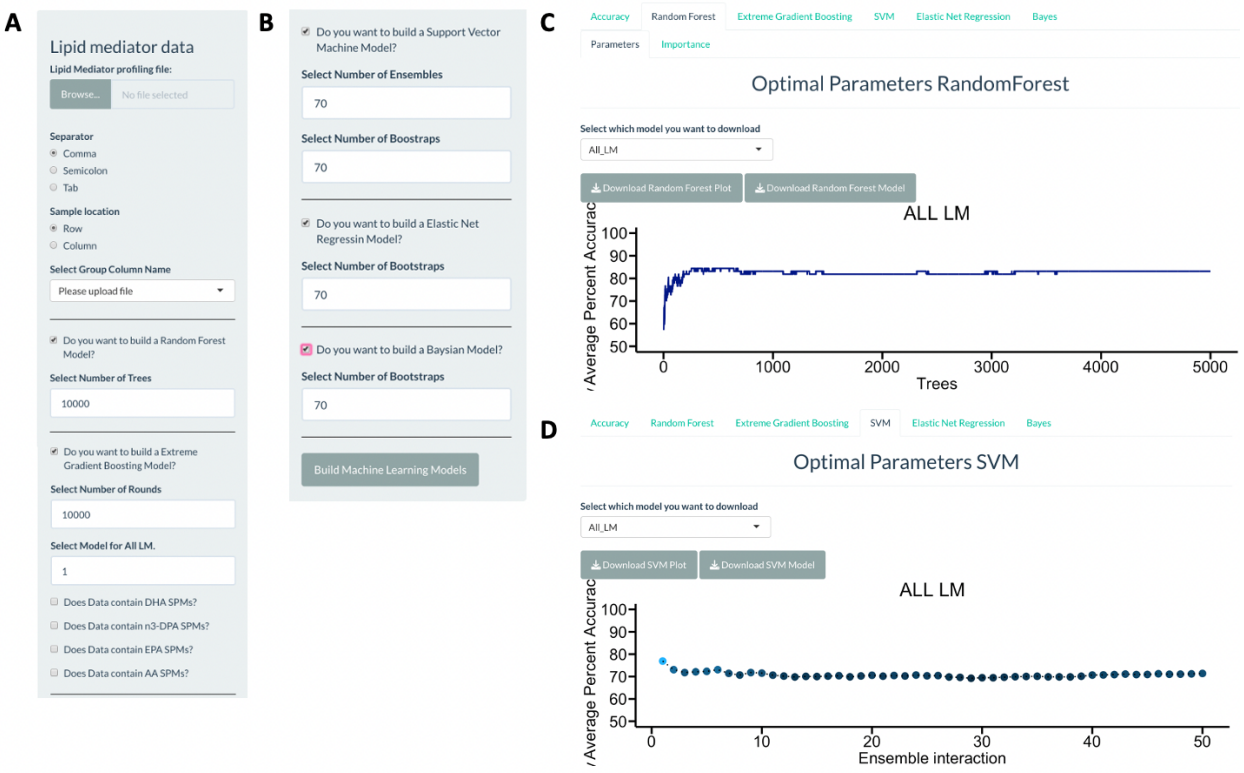


Figure 14: The sidebar panel and main panel for the Optimize ML Model page under the ML Models drop down. A) The sidebar panel with the inputs for random forest and extreme gradient boosting models. B) The inputs for SVM, Elastic Net Regression, and Bayesian GLM models. C) The random forest parameters tab with the optimal parameters plot. D) The SVM tab with the optimal parameters plot.

The outputs are the same as the Machine Learning Model page, however the models built are based on the parameters selected by the user. In figure 14C, the random forest optimal parameters plot is shown where only 5000 trees were built instead of the 10,000 trees. In addition, an SVM model was built with 50 ensembles and bootstraps instead of the default 70 ensembles and bootstraps (Figure 14D).

Build Machine Learning Model

The third page under the ML Model drop down is the Build ML Model page where the user can select which features to use to build the model. The file input and select the group name are similar to the previous pages to allow the user to upload the training dataset (Figure 15A). The file needs to contain the fatty acid family information as the first column. The third input has a list of all the mediators in the dataset and the user can select which ones they want to include in building the model (Figure 15A). The remaining inputs are a checkbox for each of the 5 machine

learning methodologies that can be used to build the model (Figure 15A). The outputs are the same as the previous two machine learning model pages however the models are built based on the selected features. The accuracy tabs outputs are shown with the % Accuracy plot, and the model summary table based on the selected mediators (Figure 15B).

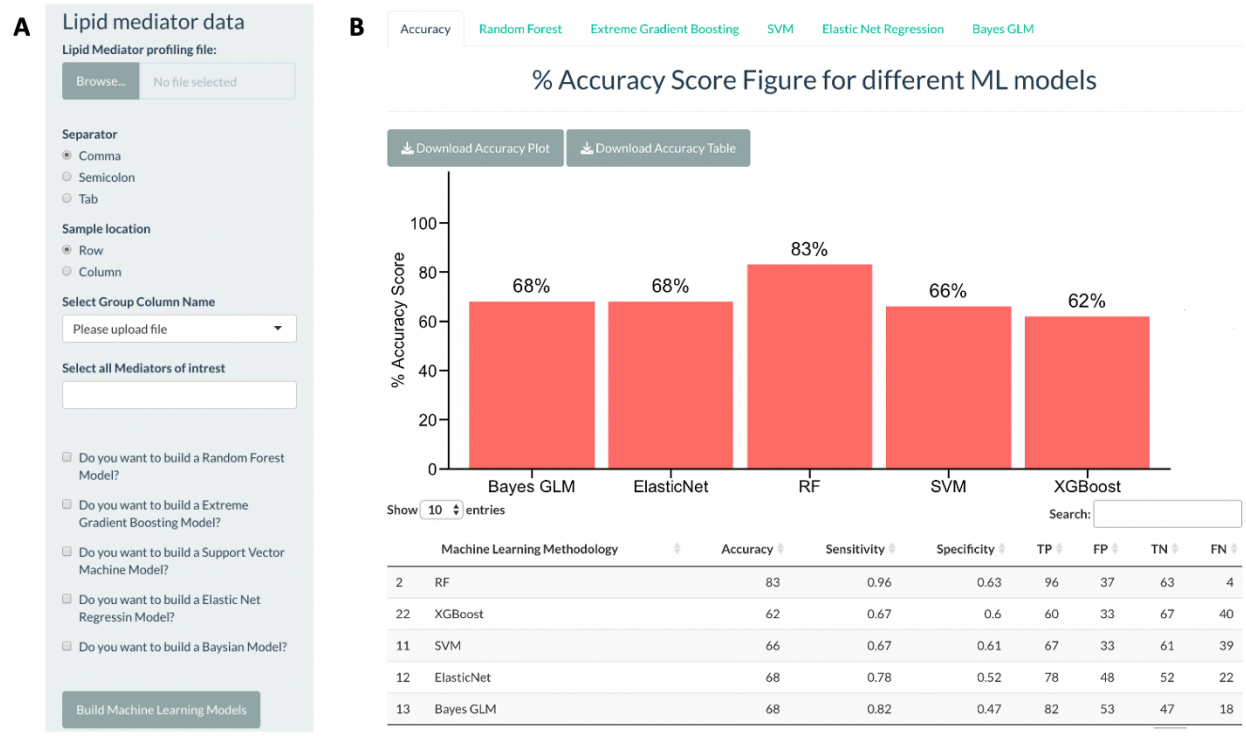


Figure 15: The sidebar panel and main panel for the Build ML Model page under the ML Models drop down. A) The sidebar panel with the inputs. B) The Accuracy tab's output of the % Accuracy plot and model summary table. for all the different Machine learning models.

Run Machine Learning Model

The final page under the ML Models drop down is the Run ML Model page where the user runs the previous models on a validation dataset. The first input is the validation dataset, followed by selecting the column with the group information (Figure 16A). Then there is a checkbox where the user can select if they would like to filter the mediators (Figure 16A). This is for models that were built using the Build Machine Learning page where not all the mediators were used to build the model. If the checkbox is selected, a list of all the mediators where the user can select the metabolites included in the model (Figure 16B). The remaining inputs are checkboxes for each machine learning methodology and the user can upload the model file

(Figure 16A). For each methodology, there is a numeric input to select how many of models of the same methodology there are, and the user can upload up to 5 models with the same methodology (Figure 14B). The first output is an accuracy tab which has all the receiver operating characteristic curve (ROC) for all the models that were run (Figure 14C). Below the plot is a table with the same table as the previous pages with an additional column for the Area Under the Curve (AUC) (Figure 14D). Then each methodology has a tab which contains the ROC plot for that model. The plots and tables can be downloaded using the download button for the page.

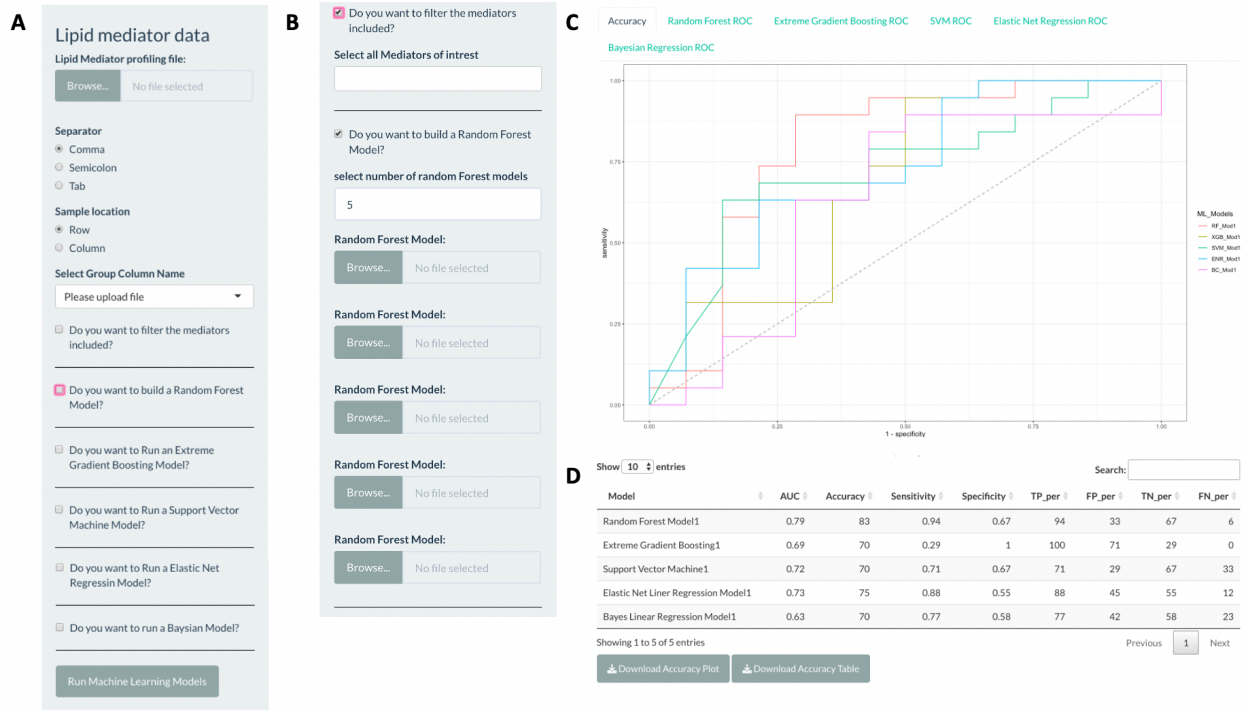


Figure 16: The sidebar panel and main panel for the Run ML Model page under the ML Models drop down. A) The sidebar panel with the inputs. B) The inputs when the user filters the mediators and uploads multiple models with the same methodology. C) The Accuracy tab's first output of the ROC curve for all the models that were run. D) The model summary table including the AUC for all the different Machine learning models.

Methodologies and Background Information

Application Building using R Shiny

The application was built using the Shiny package (Chang et al., 2021) in R which creates web application using R script. Shiny was chosen as all the previous methodologies are based on R, simplifying the integration process. Shiny scripts have two components, the user interface (ui) and the server. The ui for SPM Analyst, has been split into a sidebar where the user uploads their data and select their inputs, and a main panel where the outputs including plots and tables are displayed. The server processes the user inputs, runs the tests selected, and provides the values for the outputs which are displayed based on the ui arrangement. There are many key functions that have been used from the shiny package to build the application. This includes, fileInput() where the user can upload a file from their computer, DownloadHandler() to enable users to download tables and plots, plotOutput() to display plots, and tableOutput() to display tables.

The application was optimized to include additional features to make it more user friendly. The first is inclusion of the help text using helptext() in Shiny to add additional details for the user about the inputs and how to use the application. The second feature is error handling, which prevents the application from crashing when there is an incorrect user input. This was incorporated by using the shiny CatchApp(). Finally, the user inputs were optimized by updating the selection option to be specific to the names used in the dataset. This was done using the selectInput() and updateSelectInput(). The application was tested using datasets with different formats including comma-separated (csv) and tab separated (tsv) files to ensure it would work for those cases.

Data Processing

The data that is produced by the LC-MS analysis requires processing to account for sample loss during preparation and potential daily variations in instrument performance. The data processing steps are shown in Figure 1. The first step is to consider the potential instrument

variation by dividing the signal for reference standards obtained on the day of analysis to that obtained on the day the method was validated. The second step is to convert the area to amount in picograms using the slope of the standard curve created for each mediator with known concentrations when the method was validated. Then the sample loss during the sample preparation is considered using the normalization coefficient, which is based on the recovery percentage of the internal standards. Finally, the concentration of molecules in the sample is calculated by considering the sample weight or volume. The calculations and sub-setting of the data was completed using the base R functions. The dyplr package (Wickham et al., 2022) was used to select columns of interest and the tidyr package (Wickham and Girlich, 2022) was used to pivot the tables to be in the wider table output with mediators as columns, samples as rows, and the area or concentrations as values. The tables were rendered using Datatables (Xie, Cheng, and Tan, 2022) to produce interactive tables that the user can scroll through and search through the table.

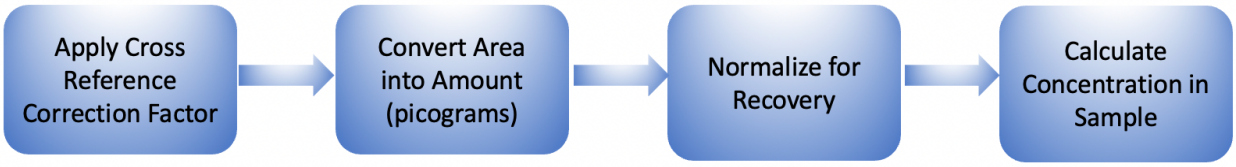


Figure 17: Flow chart of the liquid chromatography tandem mass spectrometry data processing steps.

Multivariate Statistics

PCA and PLSDA

PCA is an unsupervised dimension reduction technique, and the principal component scores are plotted for the first two or three components (Cook, Ma and Gamagedara, 2020). Although the groups may not be separated completely in the scores plot, it is useful to get a general sense of the data and identify outliers. PLS-DA analysis is a supervised dimensional reducing method that considers the group information when reducing the dataset into components (Worley and Powers, 2012). PCA and PLS-DA tests were developed using the base R and mixOmics packages (Rohart, Gautier, Singh, and Lê Cao, 2017). The PLSDA models

were evaluated using the ropls package to calculate the R2, Q2, and p-values which are calculated using permutations (Thevenoux et all., 2015). The results including the scores plot, loading plot, variance plot, and VIP scores were visualized using ggplot2 (Wickham, 2016) and plotly (Seivert, 2020) packages.

T-Test and Wilcoxon-Mann-Whitney test

The two-sample t-test is applied to each mediator to see if the means between the two groups are different (Cook, Ma and Gamagedara, 2020). The Wilcoxon-Mann-Whitney test is for data that is not normally distributed and uses a ranked set of values (Cook, Ma and Gamagedara, 2020). The two differential analysis tests are used to identify mediators have different concentrations in two groups and identify mediators of interest. For both tests, if the p-value is less than the cut off, the difference between the two groups is significant (Cook, Ma and Gamagedara, 2020). In addition, the log-fold change between the two groups is calculated and both the p-value and the log fold change are used to identify mediators of interest to be further investigated (Cook, Ma and Gamagedara, 2020). The previous methodologies for two sample t-test and Wilcoxon-Mann-Whitney test using the base R were incorporated into the application. The methodology included a normalization test to determine which test to use for the dataset. The normalization test for multivariate data included were Mardia, Henze-Zirkler, and Royston using the MVN package (Korkmaz, Goksuluk, and Zararsi, 2014). A volcano plot and pathways plots were added using ggplot2 (Wickham, 2016) and the top mediators were annotated using ggrepel (Slowikowski, 2021).

Spearman's Correlation

Spearman's correlation is a non-parametric test that does not assume frequency distribution and a linear relationship (Xiao, Ye, Esteves and Rong, 2015). It is used to determine statistically significant relationship between two variables (Xiao, Ye, Esteves and Rong, 2015). In the application, the spearman's correlation analysis can be used to relate concentrations of SPMs to other data such as gene expression or cell count. There is a previous method for spearman's correlation using the base R and the corrplot package to visualize the results (Wei

and Simko, 2021) which was incorporated in the application. This was done by generalizing the code to work for different datasets by accepting an upload file and allowing the user to specify the two groups information.

Machine Learning

Machine learning builds mathematical models based on patterns and uses those models to make decisions (Dalli, Gomez and Jouvene, 2022). Machine learning classification models include Support Vector Machine (SVM), Random Forests, Elastic Net Regression, Bayesian Generalized Linear Models (GLM), and Extreme Gradient Boosting. Machine learning models were built and used to link SPM concentrations with rheumatoid arthritis treatment response using SVM and random forest classifiers (Gomez et al., 2020). Support vector machine is based on a kernel function and performs classification of data (Cook, Ma and Gamagedara, 2020). Random Forests modeling contains an ensemble of decision trees used to classify data through binary recursive partitioning (Cook, Ma and Gamagedara, 2020). Elastic Net Regression is a logistic regression model that performs variable selection and regularization (Takahashi et al., 2020). Bayes GLM is a logistic regression based on the Bayesian inference (Rijmen 2008). Extreme gradient boosting is a method based on decision trees and is an improved version of gradient boosting machine (Stamate et al., 2019). In a study using deep learning, random forests, and extreme gradient boosting to diagnose Alzheimer's patients using metabolites in blood, extreme gradient boosting was the best predictive model (Stamate et al., 2019).

Machine learning models were built to link SPM concentrations with treatment response using SVM and random forest classifiers (Gomez et al., 2020). In addition, Bayes GLMand elastic net regression methods have been previously completed and optimized for lipid mediator profiling data. The random forest method was built using the randomForest package (Liaw and Wiener, 2002) by first determining the best number of variables randomly sampled at each tree split and using it to build the model (Cook, Ma and Gamagedara, 2020). The SVM was built using the classyfire package (Chatzimichali and Bessant, 2015) with the model set with 70 ensembles and bootstraps. The elastic net regression and Bayes GLM were built using the glmnet package (Friedman, Hastie, and Tibshirani, 2010) using 70 bootstraps. The models were

evaluated using the caret package (Kuhn, 2022) to obtain the accuracy, specificity, sensitivity, true positives, false positives, true negatives, and false negatives for each of the models.

The dataset with the rheumatoid arthritis patients was used to build a model using extreme gradient boosting to add an additional methodology for users of the application. Extreme gradient boosting is a method based on decision trees and is an improved version of gradient boosting machine (Stamate et al., 2019). In boosting machine learning methods, the trees are grown sequentially by using the information from the previous tree to build the next tree (James, Witten, Hastie and Tibshirani, 2015). The package XGBoost (Chen et all., 2022) was used to create an extreme gradient boosting machine learning model for classification of lipid mediators (Chen and Guestrin, 2016). First the data was separated into the train and test data set using the Caret package (Kuhn, 2022). Then data was converted into the matrix format and a grid with the different parameter values was built (Table 1). The hyperparameters were cross validated with the different combination of the grid and the train and test error rate were measured. The models were ranked based on the error rate and the top 50 models were built and ranked by accuracy, specificity, and sensitivity using the caret package (Kuhn, 2022). The top model was used to compare with other machine learning methods and build the importance plots. The top 5 models were used to build and iteration vs test error rate plot.

Table 1: List of Hyperparameters in the extreme gradient boosting model with their definition and values used in the optimization process (Chen et all., 2022). The values were chosen by testing the full range of values and ranking them by accuracy (Supplementary Table 1).

HYPERPARAMETER	DEFINITION	DEFAULT	VALUES
LEARNING COEFFICIENT (ETA)	The learning rate and ranges from 0-1. Prevents overfitting by making boosting process conservative.	0.3	0.01, 0.02, 0.03, 0.04, 0.05
MAXIMUM DEPTH	Maximum depth of tree.	6	1, 2, 3, 4, 6
GAMMA	The minimum loss reduction required to make a further partition on a leaf node of the tree.	0	0.5, 1, 1.5, 2, 2,5
MIN CHILD WEIGHT	Minimum sum of instance weight needed in a child.	1	1, 2, 3, 4, 5

The previous methods of SVM, random forests, bayes GLM, and elastic net regression were combined with the extreme gradient boosting method to create the machine learning page

on the application. A checkbox input was added so that only the selected machine learning model will be run. The output of each model was separated into different tabs. Since some of the models do require some time to run, a progress bar was be added with increments at every time a model was built. The same machine learning methodologies were used in the Optimize Machine Learning Model page where the model parameters can be changed. This was done using numeric inputs for the new parameters used to build the model. The 5 machine learning methods were also used in the Build Machine Learning Model page where the user can select which mediators, they want to use to build the model. The models that were built in the Machine Learning Model, Optimize Machine Learning Model, and Build Machine Learning Model pages were saved as an object using the base R saveRDS(). The final machine learning page is the Run Machine Learning Model page which takes a built model as an input and a new dataset to apply to the model. The PROC package (Robin et all., 2011) was used to calculate the receiver operator characteristic (ROC) curve plots and the area under curve (AUC) for the models.

References

Chang, W., Cheng, J., Allair, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B., 2021. shiny: Web Application Framework for R. R package version 1.7.1. https://CRAN.R-project.org/package=shiny

Chatzimichali, E., and Bessant, C., 2015. classyfire: Robust multivariate classification using highly optimised SVM ensembles. R package version 0.1-2. https://CRAN.R-project.org/package=classyfire

Cook, T., Ma, Y. and Gamagedara, S., 2020. Evaluation of statistical techniques to normalize mass spectrometry-based urinary metabolomics data. *Journal of Pharmaceutical and Biomedical Analysis*, 177, p.112854.

Dalli, J., Gomez, E. and Jouvene, C., 2022. Utility of the Specialized Pro-Resolving Mediators as Diagnostic and Prognostic Biomarkers in Disease. *Biomolecules*, 12(3), p.353.

Gomez, E., Colas, R., Souza, P., Hands, R., Lewis, M., Bessant, C., Pitzalis, C. and Dalli, J., 2020. Blood pro-resolving mediators are linked with synovial pathology and are predictive of DMARD responsiveness in rheumatoid arthritis. Nature Communications, 11(1)

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2015. *An Introduction to Statistical Learning*. New York: Springer, pp.321-323.

Korkmaz, S., Goksuluk, D., and Zararsiz, G., 2014. MVN: An R Package for Assessing Multivariate Normality. The R Journal. 6(2):151-162. https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Kuhn, M., 2022. caret: Classification and Regression Training. R package version 6.0-91. https://CRAN.R-project.org/package=caret

Rijmen, F., 2008. Bayesian networks with a logistic regression model for the conditional probabilities. International Journal of Approximate Reasoning, 48(2), pp.659-666.

Rohart, F., Gautier, B., Singh, A. and Lê Cao, K., 2017. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11), p.e1005752.

Sievert, C., 2020. Interactive Web-Based Data Visualization with R, plotly, and shiny. Chapman and Hall/CRC Florida. https://plotly-r.com

Slowikowski, K., 2021. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://CRAN.R-project.org/package=ggrepel

Thevenot EA, Roux A, Xu Y, Ezan E, Junot C (2015). "Analysis of the human adult urinary metabolome variations with age, body mass index and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses." *Journal of Proteome Research*, **14**, 3322-3335.

Wei, T., and Simko, V., 2021. R package 'corrplot': Visualization of a Correlation Matrix (Version 0.92). https://github.com/taiyun/corrplot

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., and Müller, K., 2022. dplyr: A Grammar of Data Manipulation. R package version 1.0.8. https://CRAN.R-project.org/package=dplyr

Wickham, H., and Girlich, M., 2022. tidyr: Tidy Messy Data. R package version 1.2.0.

Worley, B. and Powers, R., 2012. Multivariate Analysis in Metabolomics. *Current Metabolomics*, 1(1), pp.92-107.

Xiao, C., Ye, J., Esteves, R. and Rong, C., 2015. Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, 28(14), pp.3866-3878.

Xie, Y., Cheng, J., and Tan, X., 2022. DT: A Wrapper of the Library 'DataTables'. R package version 0.22. https://CRAN.R-project.org/package=DT