**ECCO Functionality**

**Create Clustergrams:**

The create clustergrams functionality allows users to create clustergrams, using any of the four common linkage functions (single, ward, average, complete), and broad range of distance metrics. Note that the Ward will only allow the user to select the Euclidean distance metric. Prior to analysis the user is allowed to select the transformation, data scaling and color-map they want before being prompted to input the data in which they would like to have clustered.

A screenshot of a computer

Description automatically generated with medium confidenceInput: Data should be input as an **\*.xlsx** file, groups can be specified by matching inputs to the second row of the input data sheet. **Note:** **All data sheets should contain an m/z column and a rtmed column.** If the input data does not contain m/z or rtmed data, simply inputting 1 – N (where N is number of rows of data) in the m/z column, and zeros in the rtmed column will ensure that your data is appropriately analyzed. Up to 24 different groups can be specified.

**Groups**

A picture containing chart

Description automatically generated

Options: Upon clicking the create clustergram button the user will be prompted with to first select a linkage function (e.g., ward, select by double clicking), the user will then be able to select the distance metric, color-map, data transform, and data scaling, respectively. Before submitting for creation, of the clustergram.

Chart, treemap chart

Description automatically generatedExample Output:

**Compare linkage functions:**

The compare linkage functions functionality is aimed at allowing the user to compare various linkage function’s ability to cluster a given dataset. One to four linkage functions can be analyzed using this functionality.

Input: Datasheets should contain the data objects of interest in rows and the subsequent features in the columns, the data should be skirted by an m/z and a rtmed column, see create clustergram input if data does not have these values.

A screenshot of a computer

Description automatically generated with medium confidence

Options: Upon clicking the compare linkage functions button the user will first need to select the distance metric of interest, followed by selection of the number of comparisons, linkage functions, data transform and finally the data scaling.

Shape

Description automatically generated with low confidenceExample Output:

**Cluster optimization:**

The cluster validation functionality allows the user to select one of five different clustering validation metrics (see methods section). This functionality uses a minimum spanning tree approach as the underlying clustering of data that is validated.

Input: The input data should mimic that of the input data for comparison of linkage functions. A screenshot of a computer

Description automatically generated with medium confidence

Options: Upon selection the validation metric the user will be prompted to select the data transform, followed by the data scaling, and validation technique.

**Text

Description automatically generated**Example Output:

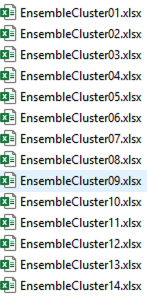
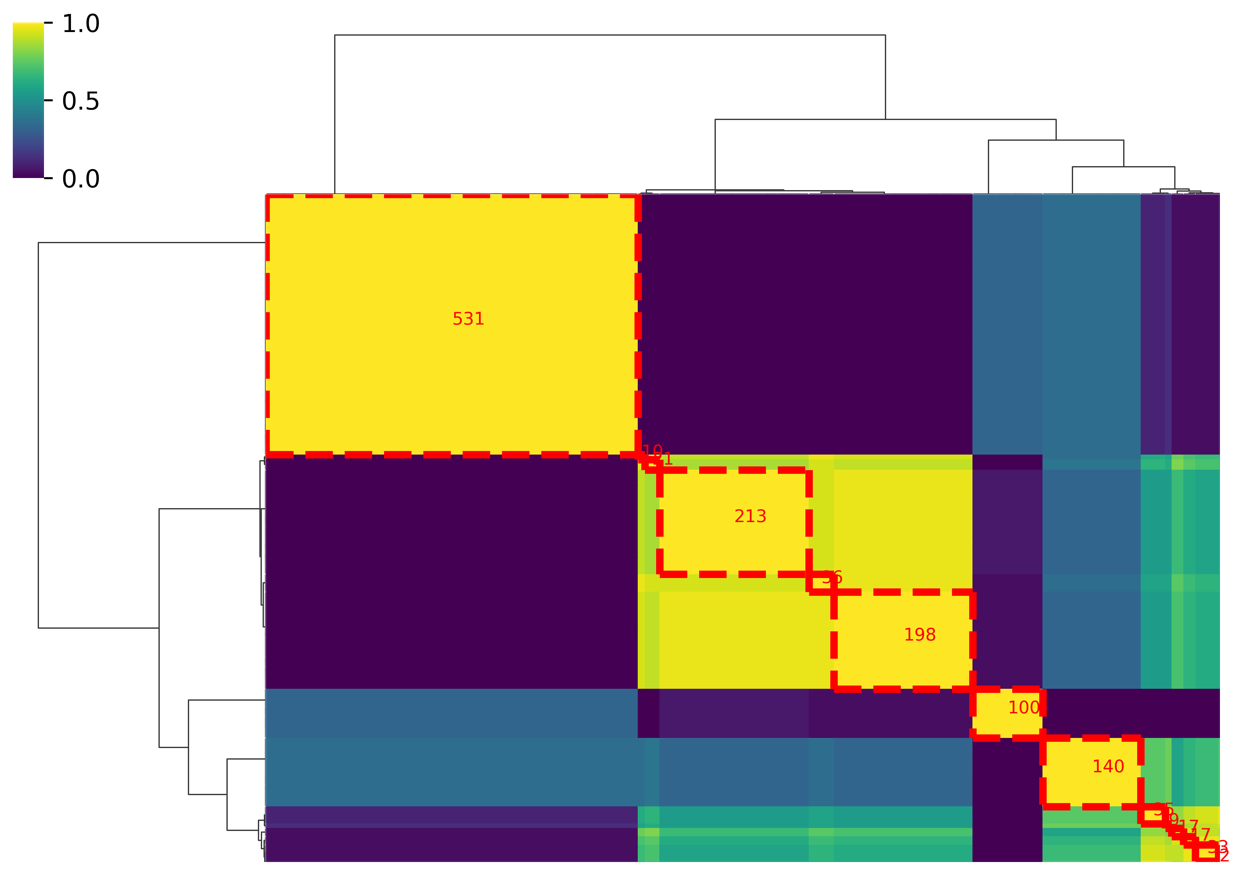
**Ensemble Clustering:**

The ensemble clustering functionality allows the user to an ensemble size of interest by selecting the linkage functions they would prefer to use followed the distance metrics of interest and finally the user is prompted to select the color-map, data-transform, data-scaling, the number of clusters (i.e., the appropriate number of clusters from cluster validation), and the number of data objects for a cluster to be considered relevant. Additionally, should the user forget to run a cluster validation ahead of selecting the ensemble clustering step there is an option to run a cluster validation that will feed directly into the stand ensemble size presented in this paper. Additionally, should users be interested in utilizing the ensemble proposed in this paper a “Standard Ensemble” button is provided on the first interface allowing the user to skip the steps of selecting the linkage functions, and distance metrics of interest.

A screenshot of a computer

Description automatically generated with medium confidenceInput: The input data should mimic that of the input data for comparison of linkage functions.

Example Output: Where the dashed-red boxes represent the clusters in which the data objects clustered together across all 31 of the clustering solutions, these are automatically output excel sheets.



**Peaks to Pathways:**

The peaks to pathways functionality allows users to input the clusters recommended through ensemble clustering analysis (i.e., the clusters outlined in dashed red-lines) for the creation of “peaks-to-pathways” files that can be input to the mummichog algorithm (cite).

A screenshot of a computer

Description automatically generated with medium confidenceInput: The first input data should mimic that of the input data for comparison of linkage functions, the second selection should be a directory containing the recommended ensemble clustering files.

See output from ensemble clustering functionality, for second input reference.

Table

Description automatically generated with medium confidenceOutput: Peaks to pathways files containing columns for m/z, p-value, and retention times. P-values are set to 0.04 to signify significance under the standard 0.05 threshold for data objects in which are found within a cluster, all other data objects detected in the metabolomic analysis are given a p-value of 1. Each output file will have the structure below.

**Group Medians:**

The group medians functionality allows users to input a file in which all samples from a given experiment were analyzed, for the group medians to be calculated. This allows the users to understand automatically generate group medians of interest for analysis.

A picture containing table

Description automatically generatedInput: The input file should be the same format as that submitted to Metaboanalyst (cite).

Example Output:

Graphical user interface, application, table, Excel

Description automatically generated

At this point the retention times must be added to the far right column manually.

**Data Integrity:**

The data integrity functionality allows users to correct for the introduction of double decimals to the data following analysis in Metaboanalyst (cite).

Table

Description automatically generatedInput: A datasheet containing m/z values with more than one decimal.

Output: An excel workbook with all the double decimals removed.

**Cluster Selection:**

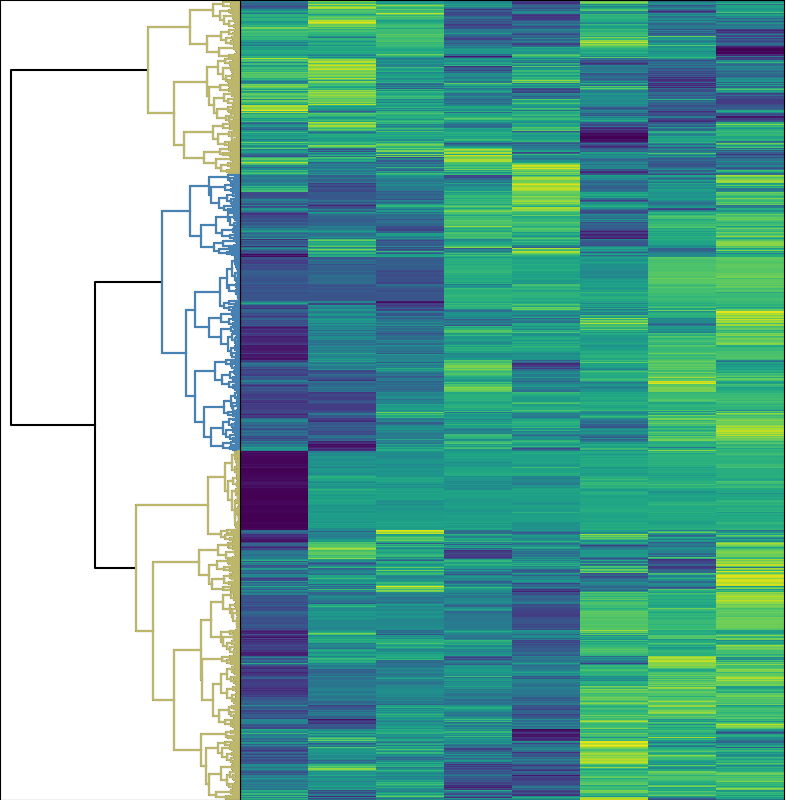
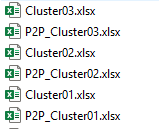
The cluster selection functionality allows the user to input a set of data, such as that submitted to the create a clustergram function, and then select the clusters of interest.

A screenshot of a computer

Description automatically generated with medium confidenceInput: See linkage comparison input section for full details.

Options: The user is prompted to make the same selections as when creating a clustergram (i.e., color-map, data scaling and transformation, linkage function, and distance metric).

Example Output:



**Selected Clusters Figure:**

The selected clusters figure functionality is aimed at providing the users with a publication quality pdf of the clusters which were selected using the cluster selection tool

Input: The user needs to select the Heatmap.xlsx file generated during the cluster selection

Example Output:  (see next page)

