**Global\_Configs**

* *‘data\_path’:* add path to folder where all input files will be stored
* *‘result\_path’:* add path to folder where all result data should be stored

**01\_Pre\_Processing\_SC\_Data**

Only needs to be specified/ adjusted if sc data should be used

* *'data\_name':* enter name of the sc dataset (Default: Prepared\_sc\_Data); the dataset needs to be of format 'h5seurat' and contain two meta-data columns: 'cluster\_id' [specifies the cell-type, cluster for aggregation to pseudobulk]; 'sample\_id' [sample identifier; needs to be the same across all integrated datasets]
* *'data\_type':* 'h5seurat'

**02\_Pre\_Processing\_Configs\_SC**

Only needs to be specified/ adjusted if sc data should be used

* *'configuration\_name'* :

*‘Values*: any text

*‘Description’*: enter a name that will be added to all resulting tables and figures of this execution

* *'data\_name':* enter name of the sc dataset (Default: Prepared\_sc\_Data); the dataset needs to be of format 'h5seurat' and contain two meta-data columns: 'cluster\_id' [specifies the cell-type, cluster for aggregation to pseudobulk]; 'sample\_id' [sample identifier; needs to be the same across all integrated datasets]
* *'data\_type':* 'h5seurat'
* *'cell\_expr\_thres1':*

‘Values’: two numbers separated by ‘;’ (‘x;y’ - Default: ’50;10’)

‘Description’: defines threshold for filtering of genes

* first number: percentage value (x%) - gene needs to be expressed in at least x% of cells to be used as input,
* second number: average minimum amount of cells per sample - gene needs to be expressed in at least y \* number of samples cells

Both thresholds need to apply

*'cell\_expr\_thres2'*: same as cell\_expr\_thres1; different numbers may be entered; gene will be used in analysis if either thres1 **or** thres2 applies (Default: ’40;20’)

* *'cell\_type\_exclusion'*:

*‘Values*’: enter names from the ‘cluster\_id’ column of the sc-data file separated by ‘,’

*‘Description’:* specify cell-types that should be excluded from further analysis (e.g. because of low amount of cells); names need to match exactly to values in 'cluster\_id' information; values need to be entered comma seperated without space in between

**02\_Pre\_Processing\_Configs**

Specify one row for each dataset that should be included

* *'configuration\_name'*:

*Values: any text*

*Description:* enter a name that will be added to all resulting tables and figures of this (should be the same as in previous config file)

* *'data\_name':*

*Values:* enter names of the datasets which should be used in the MOFA analysis; all datasets listed here will be integrated (if single-cell data is used also set up the previous configuration file: ‘02\_Pre\_Processing\_Configs\_SC’)

* *'data\_type':*

*Values:*  'h5seurat' or ‘csv’

* *‘remove\_sample\_ids’:*

*Values: ‘*sample\_id’ s separated by ‘,’ or empty

*Description:* specify in case samples should be removed from the analysis

* *‘sample\_filtering\_thres’*

*Values:* number between 0 and 1 (x)

*Description:* filters out samples for which across all features (e.g. genes) more than x% of the values are zero *(Usage filter out low quality samples)*

* *‘feature\_filtering\_thres’:*

*Values:* number between 0 and 1 (x)

*Description:* specifies whether features should be filtered out that are not expressed in a x% of samples [0: no features will be filtered; 1: all features will be filtered 🡪 error ; default: 0.2]

* *‘library\_adjustment’:*

*Values:* either ‘TRUE’ or ‘FALSE’

*Description:* specifies whether library size adjustment will be applied to data, meaning that all samples will have the same amount of counts across all features after the transformation [usually applied to single-cell and bulk RNA-seq data]

* *‘log\_transformation’:*

*Values:* either ‘TRUE’ or ‘FALSE’

*Description:* specify whether values should be log-transformed

* *‘variable\_gene\_filtering’:*

*Values:* between 0 and 1 (x)

*Description:* only the x% of the most variable genes (based on their variance across samples) will be kept

* *‘quantile\_normalization\_samples’:*

*Values:* either ‘TRUE’ or ‘FALSE’

*Description:* specify whether quantile normalization across samples should be applied [in our data applied to single-cell and bulk RNA-seq data]

* *‘ribosomal\_mitochondrial\_gene\_filtering’:*

*‘Values’:* either ‚TRUE‘ or ‘FALSE’

*Description:* specify whether ribosomal and mitochondrial genes should be excluded from analysis [note: this only works when gene names are given by their ‘SYMBOL’ annotation]

* *‘feature\_wise\_quantile\_normalization’:*

*Values:* either ‚TRUE‘ or ‘FALSE’ (needs to be the same for all entries)

*Description:* specifies whether feature-wise quantile normalization should be applied to all features or not

**03\_MOFA\_Configs:**

Specify one row for each different mofa configuration that should be run

* *'configuration\_name'*:

*Values:* any text

*Description:* enter the name of the configuration by which data in previous steps has been generated (‘02\_Preprocessing\_Configs.csv’)

* *‘mofa\_result\_name’:*

*Values:* any text

*Description:* enter a name that will be added to all resulting tables and figures that are generated as result of the MOFA model

* *‘amount\_of\_factors’*

*Values: numeric value* (max: amount of features -1; min: 1; default: 20)

*Description:* amount of latent factors that should be estimated by the MOFA model

* *‘weighting\_of\_views’*

*Values: ‘TRUE’ , ‘FALSE’*

*Description:* defines whether the views should be weighted based on the amount of features (if ‘TRUE’, views with a lower amount of features will receive a higher weight)

* *‘scale\_views’*

*Values: ‘TRUE’ , ‘FALSE’*

*Description:* defines whether the features should be scaled based on the corresponding MOFA functionality

**04\_Factor\_Analysis\_Configs:**

* *'configuration\_name'*:

*Values:* any text

*Description:* enter the name of the configuration by which data in previous steps has been generated and was used as input for the MOFA model (‘02\_Preprocessing\_Configs.csv’)

* *‘mofa\_result\_name’:*

*Values:* any text

*Description:* enter the name of the MOFA results output for which the analysis should be done (name used in: ‘03\_MOFA\_Configs.csv’)

* *‘relevant\_factors’:*

*Values*: text, factor names that should be plotted comma separated (e.g.: Factor1,Factor2,Factor3)

*Description:* subselection of the factors of the MOFA model that will be plotted (Default: ‘Factor1,Factor2,Factor3,Factor4,Factor5’)

* *‘numeric covariates’:*

*Values:* text, name of the numeric covariates that should be analyzed comma separated (e.g. Age,Weight); needs to correspond to columns in Sample\_Meta\_Data.csv)

*Description:* used to generate correlation plots between the specified covariates + factors

* *‘categorical\_covariates’:*

*Values:* text, name of the categorical covariates that should be analyzed comma separated (e.g. Gender,Disease); needs to correspond to columns in Sample\_Meta\_Data.csv)

*Description:* used to generate boxplots to analyze differences in factor values

* *‘top\_variable\_thres’:*

*Value:* numeric value between 0 and 1 (Default: 0.005)

*Description:* defines the selection of top features per factor that should be analyzed

**05\_Feature\_Analysis\_Configs:**

* *'configuration\_name'*:

*Values:* any text

*Description:* enter the name of the configuration by which data in previous steps has been generated and was used as input for the MOFA model that should be analyzed (‘02\_Preprocessing\_Configs.csv’)

* *‘mofa\_result\_name’:*

*Values:* any text

*Description:* enter the name of the MOFA results output for which the analysis should be done (name used in: ‘03\_MOFA\_Configs.csv’)

* *‘factor’:*

*Values:* factor name (e.g. Factor1, Factor2; Default: ‘Factor2’)

*Description:* select the MOFA Factor for which top features should be plotted in the heatmap

* *‘top\_variable\_thres’:*

*Values:* numeric value between 0 and 1 (x) (Default: 0.005)

*Description:* select the x% of top features (meaning having the highest weights) that should be displayed in heatmap (! if value is to large heatmap might include to many features and not be readable anymore)

* *‘faceting\_variable’:*

*Values:* name of a column in the ‘sample\_meta\_data.csv’ (needs to be a categorical column)

*Description:* faceting/ grouping of samples in the heatmap is done based on this parameter

* *‘type’:*

*Values:* name of a type/view for which features should be selected; if empty all views with top features will be displayed

*Description:* defines of which view features will be displayed in the heatmap

**06\_Pathway\_Configs:**

* *‘mofa\_result\_name’:*

*Values:* any text

*Description:* enter the name of the MOFA results output for which the analysis should be done (name used in: ‘03\_MOFA\_Configs.csv’)

* *‘factor\_set’:*

*Values:* numbers comma separated (e.g: ‘1,2,3’)

*Description:* defines the factors for which the enrichment analysis should be conducted by default for factors 1-5 (‘1,2,3,4,5’)

* *‘coverage\_par’:*

*Values:* numeric between 0 and 1 (x)

*Description:* within the MOFA feature set at least x% of the genes of the pathway need to be included in order to test the pathway for enrichment, i.e. pathways with a lot of genes that are not in the feature set will not be tested

* *‘types’:*

*Values:* name of a type/view for which the enrichment analysis should be executed

*Description:* for all names entered here a view-specific enrichment analysis will be executed

* *‘coverage\_plot’:*

*Values: :* numeric between 0 and 1 (x)

*Description:* after the execution of the pathway enrichment analysis a plot will be generated showing selected enriched pathways and the corresponding genes; this paremeter is used to filter the pathways shown in the plot and indicates to show only pathways for which x% of genes have been included in the MOFA feature set

* *‘p\_value\_plot’:*

*Values:* numeric between 0 and 1 (x)

*Description:* after the execution of the pathway enrichment analysis a plot will be generated showing selected enriched pathways and the corresponding genes; this paremeter is used to filter the pathways shown in the plot and indicates to show only pathways for which the p-value of the enrichment is smaller than x

* *‘enrichment\_plot:*

*Values:* either ‘positive’, ‘negative’ , ‘complete’

*Description:* pathway enrichment is executed in three directions: analyzing only features with positive weights, only negative weights and absolute values (‘complete’); this parameter is used to filter the pathways shown in the plot and indicates to show only pathways of a certain type of enrichment (‘positive’, ‘negative’ , ‘complete’)

* *‘top\_features\_plot’:*

*Values:* numeric between 0 and 1 (x)

*Description:* this parameter is used to filter the genes shown in the plot after the pathway enrichment analysis, so only features among the top x% of features will be shown in the plot

**07\_Comparison\_Configs:**

* *‘mofa\_result\_name’:*

*Values:* any text

*Description:* enter the name of the MOFA results output for which the comparisons should be done (name used in: ‘03\_MOFA\_Configs.csv’); add a new entry for each mofa result that should be compared

* *‘compare\_factors’:*

*Values:* factor names separated by ‘,’ (e.g. ‘Factor1,Factor2,Factor3’)

*Description:* the different MOFA models will be compared for the factors specified here (Default: ‘Factor1,Factor2,Factor3,Factor4,Factor5)