1) Preprocessing

normFunc(x)
Inputs
x: A vector of numbers to convert to z-scores. Generally each cell or patient sample would use this conversion in the DEGAS workflow.
Output
A vector of z-scores based on the original numeric vector

scaleFunc(x)
Inputs
x: A vector of number to scale to [0,1]. Generally each cell or patient sample would use this conversion after conversion to z-scores using normFunc.
Output
A vector of [0,1] scaled values

initDEGAS()
This function must be run to initialize the DEGAS parameters before training the model. The following functions can be used to change the initialized DEGAS parameters to the user's preference.
Inputs
None
Output
None

setPython(path2python)
Inputs
path2python: A string containing the path to the location of the preferred python executable. This is important if the tensorflow package is only available for a specific version of python on the user's computer.
Output
None

set_training_steps(inp)
Inputs
inp: The number of training steps during DEGAS model training.
Output
None

set_single_cell_batch_size(inp)
Inputs
inp: The number of cells to use in the minibatch
Output
None

set_patient_batch_size(inp)
Inputs
inp: The number of patients to use in the minibatch
Output
None

set_hidden_feature_number(inp)
Inputs
inp: The number of features in each hidden layer of the DEGAS model.
Output
None

set_dropout_keep_fraction(inp)
Inputs
inp: The percentage of nodes to keep during dropout.
Output
None

set_l2_regularization_term(inp)
Inputs
inp: The regularization term weight lambda for L2 regularization.
Output
None

set_patient_loss_term(inp)
Inputs
inp: The term weight lambda for the patient label loss.
Output
None

set_MMD_loss_term(inp)
Inputs
inp: The term weight lambda for the MMD loss.
Output
None

2) Model training and prediction

runCCMTLBag(scExp, scLab, patExp, patLab, tmpDir, model_type, architecture, FFdepth, Bagdepth)
Inputs
scExp: A matrix (rows=cells, columns=genes) of expression values from scRNA-seq data. The columns (i.e. genes) should be in the same order as the patExp matrix.
scLab: A matrix (rows=cells, columns=labels) of binary cell labels corresponding to the scExp cells. This is used in the case where cell type or cell cluster labels exist for the scRNA-seq data. Each row is onehot encoded such that each row contains a one in a single column and zeros in the rest of the columns. In the case that no such labels exist NULL should be passed to this argument.
patExp: A matrix (rows=patient samples, columns=genes) of expression values from bulk expression data. The columns (i.e. genes) should be in the same order as the scExp matrix.
patLab: A matrix (rows=patient samples, columns=labels) of patient sample labels corresponding to the patExp samples. In the case of classification tasks there should exist two or more columns where each row contains a single one with all other columns in that row containing a zero. In the case of survival tasks there should be two columns. The first column

will contain times and the second column should contain event status (1=event, 0=censored). In the case that no such labels exist NULL should be passed to this argument.
tmpDir: This string argument indicates the path to the tmp directory used for processing (e.g. "~/Desktop/tmp/")
model_type: This string argument indicates the type of model used. Options include "BlankClass", "ClassBlank", "ClassClass", "BlankCox", and "ClassCox". This indication should match the input matrices for scLab and patLab. As an example, for a BlankClass model, scLab should be NULL and patLab should be a onehot matrix of patient sample labels.
architecture: This string argument can either be "Standard" or "DenseNet". This specification indicates which type of model to train where "Standard" is a feed forward network and DenseNet is a dense net network.
FFdepth: This numeric argument indicates the number of layers in the model. This number must be greater than or equal to 1. Please note that very large numbers may take long times to run or use all of the available RAM. Also, note that Standard and DenseNet models are the same if this argument is set to 1.
Bagdepth: This numeric argument indicates the number of times to bootstrap aggregate the models. This argument must be greater than or equal to 1. Please not that if equal to 1, there is no bootstrap aggregation.
Output
A trained bootstrap aggregated DEGAS model.

predClassBag(ccModel, Exp, scORpat)
Inputs
ccModel: This is a bootstrap aggregated DEGAS model that was previously trained with the runCCMTLBag function.
Exp: This is a matrix (rows=cells/samples, columns=genes) of expression values from which labels will be predicted. The order of the column genes should be identical for the prediction dataset and the training dataset.
scORpat: This is a string ("sc" or "pat") to tell the algorithms whether to output the predicted patient sample label ("pat") or the predicted cellular label ("sc"). It is worth noting that for BlankClass and BlankCox models the cellular label can not be predicted. For ClassBlank models the patient sample label cannot be predicted.
Output
A matrix of label probabilities based on the new input data, the trained bootstrap aggregated DEGAS model, and the output label of choice.

runCCMTL(scExp, scLab, patExp, patLab, tmpDir, model_type, architecture, FFdepth)
See runCCMTLBag. The only difference is that this function does not train a bootstrap aggregated model. The inputs and outputs are the same as runCCMTLBag except that the model is not bootstrap aggregated.

predClass(ccModel1, Exp, scORpat)
See predClassBag. The only difference is that it predicts outcomes from a singular DEGAS model, i.e. not bootstrap aggregated. The inputs and outputs are the same as predClassBag except that the model is not bootstrap aggregated.

3) Post processing

knnSmooth(probs, locs, k)
Inputs

probs: A numeric matrix (rows=cells/samples, columns=labels) which will be smoothed. For the purposes of the general DEGAS workflow, this will be the association matrix which can be smoothed to remove some of the random noise in the labels.
locs: A numeric matrix of locations for the row in the probs matrix. For the purposes of the general DEGAS workflow, this will generally be PCA/tSNE/UMAP coordinates of the sample association matrix.
k: A number indicating the number of neighbors to smooth with. Default is 5.

Output

A numeric matrix with the same dimensions as probs that has been smoothed based on it's neighbors.

toCorrCoeff(probs)

Inputs

probs: A numeric matrix (rows=cells/samples, columns=labels) which will be smoothed. For the purposes of the general DEGAS workflow, this will be the association matrix which can be smoothed to remove some of the random noise in the labels.

Output

A numeric matrix where the probability matrix values [0,1] have been converted to associations [-1,1].