**Explanation of Data**

About the data file name:

A larger rmFactor means more PDM samples are removed as outliers. rmFactor 0.0 means no PDM samples are removed as outliers.

ddNorm indicates the approach used to normalize drug descriptors. ‘std’ is regular standardization. ‘quantile’ is quantile normalization.

res is a data frame of drug response data:

SOURCE: study names, including CCLE, CTRP, GDSC, gCSI, NCI60, PDM

ccl\_name: CCL IDs.

ctrpDrugID: drug IDs.

area\_under\_curve: drug responses. For CCL studies, area\_under\_curve is the AUC value calculated by Fangfang. For PDM, area\_under\_curve is log2 transformation of RM-EFS.

groupID: it is not meaningful for CCL studies, but in PDM study it is the unique ID for the combination of a PDM lineage and a drug. There are 599 different groupIDs ranging from 1 to 599.

drug is a data frame of drug descriptor and fingerprint features.

Columns are features. In the columns names, ‘DD’ indicates drug descriptor, ‘ECFP’ indicates ECFP fingerprint, ‘PFP’ indicates PFP fingerprint, ‘naLabeled’ indicates binary variable labeling missing drug descriptor values, ‘num’ after ‘|’ indicates numeric variable, ‘int’ after ‘|’ indicates binary variable.

genomics is a data frame of gene expression data. Columns are features.

The first ~2000 features are individual genes, indicated by ‘geneGE’ in column name. The other features are pathway level gene expression features indicated by ‘c2cp’ in column name.

**Explanation of PDM\_10Fold\_Partition**

This folder includes partition files of 10 10-fold cross-validations (totally 100 cross-validation trials). For each cross-validation trail, there are three files, which are TrainList.txt, ValList.txt, and TestList.txt. These files include the groupIDs of combinations of PDM lineages and drugs.

**Explanation of CCL\_10Fold\_Partition**

This folder includes partition files of CCLs in CCL studies. For each study, there are 10 cross-validation trails. For each cross-validation trail, there are three files, which are TrainList.txt, ValList.txt, and TestList.txt. These files include the CCL IDs used in partition. We always do hard partition for CCLs but not for drugs in cross-validation on CCL data.

**Explanation of ExampleCode**

Example\_LoadData.py Example code for loading the data.

Example\_TransferLearningOnPDM\_FixedCV.py Code used for transfer learning from CCL to PDM using lightGBM. It includes code for transformation of PDM drug response value for transfer learning. The transformation includes (1) log2 transformation (2) subtraction from the maximum value (3) standardization to have the same mean and standard deviation as the CCL drug response data.

Example\_ModelTransferAnalysisOnCCL.py Code for model transfer analysis on CCL data using lightGBM.

**Performance Measures**

We need to calculate the following performance measures for the analysis.

1. R2
2. Mean absolute error
3. Spearman rank correlation coefficient
4. Log10 transformation of p-value of correlation coefficient.

R2 will be used as the main performance measure to select hyper parameters in model training on CCL data.

When calculating prediction performance on PDM data, prediction outcomes of multiple samples associated with the same groupID need to be averaged and true drug responses associated with the same groupID need to be averaged. Then, performance measures on PDM data can be calculated.

**Analysis Scheme**

1. CCL -> PDM transfer learning

Step 1: 10-fold cross-validation on a CCL dataset with a hard partition of CCLs. Multiple hyperparameter settings will be tested and the best hyperparameter setting will be identified based on testing performance of R2. The cross-validation for selecting hyperparameters needs only training set and testing set, so the training list and the validation list in partition files can be combined for training and the testing list can be used for testing.

Step 2: Train a prediction model on the whole CCL dataset using the best hyperparameter setting from Step 1.

Step 3: Transformation of PDM drug response values to match the distribution of CCL drug response values. Log2 transformation of PDM drug response values has already been done. You will need to do (1) subtraction from the maximum value for inverting and (2) standardization to have the same mean and standard deviation of the CCL drug response data.

Step 4: Transfer learning on the PDM data. The model generated in Step 2 will be refined and tested on PDM for 10 10-fold cross-validations (totally 100 trials), i.e. model refinement and testing for 100 times.

Step 5: Calculate prediction performances. (1) In each 10-fold cross-validation (10 trails), each PDM sample will have one prediction outcome. (2) Take an average of prediction outcomes of multiple samples associated with the same groupID. (3) Take an average of true drug responses associated with the same groupID. (4) Calculate performance measures based on the average values. (5) After doing (1)-(4) for 10 10-fold cross-validations, calculate the mean and standard deviation of each performance measure.

1. PDM only analysis

Step 1: Do 10 10-fold cross-validations (100 trails) on PDM data. No additional transformation of PDM drug response values is needed.

Step 2: Calculate prediction performances. (1) In each 10-fold cross-validation (10 trails), each PDM sample will have one prediction outcome. (2) Take an average of prediction outcomes of multiple samples associated with the same groupID. (3) Take an average of true drug responses associated with the same groupID. (4) Calculate performance measures based on the average values. (5) After doing (1)-(4) for 10 10-fold cross-validations, calculate the mean and standard deviation of each performance measure.

1. CCL -> CCL transfer learning

Step 1: 10-fold cross-validation on the source CCL dataset with a hard partition of CCLs. Multiple hyperparameter settings will be tested and the best hyperparameter setting will be identified based on testing performance of R2. The cross-validation for selecting hyperparameters needs only training set and testing set, so the training list and the validation list in partition files can be combined for training and the testing list can be used for testing.

Step 2: Train a prediction model on the whole source CCL dataset using the best hyperparameter setting from Step 1.

Step 3: Transfer learning on the target CCL dataset. The model generated in Step 2 will be refined and tested in 10-fold cross-validation. 10 model refinements and tests.

Step 4: All four performance measures will be calculated.