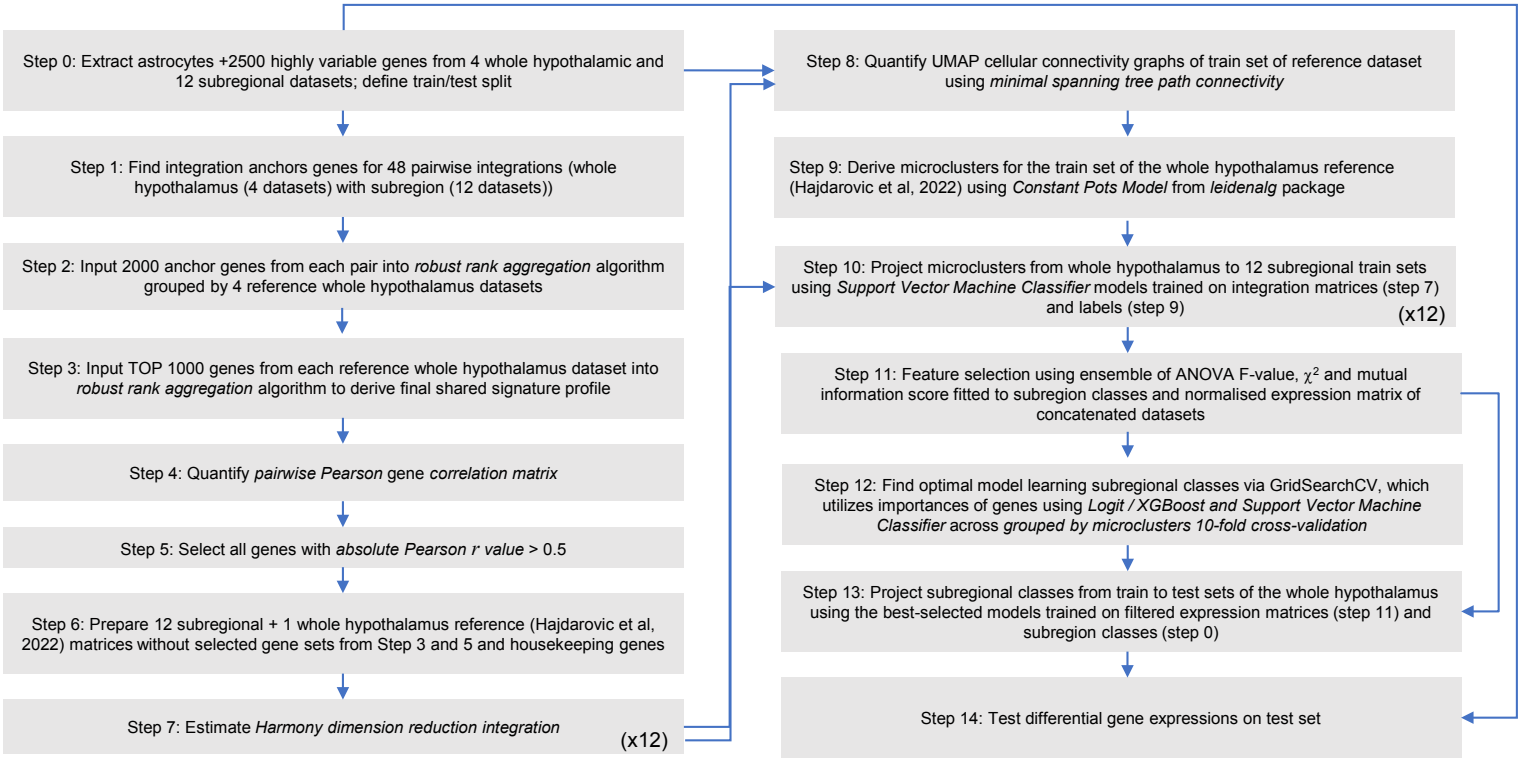
Step 0: Extract astrocytes +2500 highly variable genes from 4 whole hypothalamic and 12 subregional datasets; define train/test split
Step 1: Find integration anchors genes for 48 pairwise integrations (whole hypothalamus (4 datasets) with subregion (12 datasets))
Step 2: Input 2000 anchor genes from each pair into *robust rank aggregation* algorithm grouped by 4 reference whole hypothalamus datasets
Step 3: Input TOP 1000 genes from each reference whole hypothalamus dataset into *robust rank aggregation* algorithm to derive final shared signature profile
Step 4: Quantify *pairwise Pearson* gene *correlation matrix*
Step 5: Select all genes with *absolute Pearson r value* > 0.5
Step 6: Prepare 12 subregional + 1 whole hypothalamus reference (Hajdarovic et al, 2022) matrices without selected gene sets from Step 3 and 5 and housekeeping genes
Step 7: Estimate *Harmony dimension reduction integration*
(x12)
Step 8: Quantify UMAP cellular connectivity graphs of train set of reference dataset using *minimal spanning tree path connectivity*
Step 9: Derive microclusters for the train set of the whole hypothalamus reference (Hajdarovic et al, 2022) using *Constant Pots Model* from *leidenalg* package
Step 10: Project microclusters from whole hypothalamus to 12 subregional train sets using *Support Vector Machine Classifier* models trained on integration matrices (step 7) and labels (step 9)
(x12)
Step 11: Feature selection using ensemble of ANOVA F-value, χ^2 and mutual information score fitted to subregion classes and normalised expression matrix of concatenated datasets
Step 12: Find optimal model learning subregional classes via GridSearchCV, which utilizes importances of genes using *Logit / XGBoost and Support Vector Machine Classifier* across *grouped by microclusters 10-fold cross-validation*
Step 13: Project subregional classes from train to test sets of the whole hypothalamus using the best-selected models trained on filtered expression matrices (step 11) and subregion classes (step 0)
Step 14: Test differential gene expressions on test set