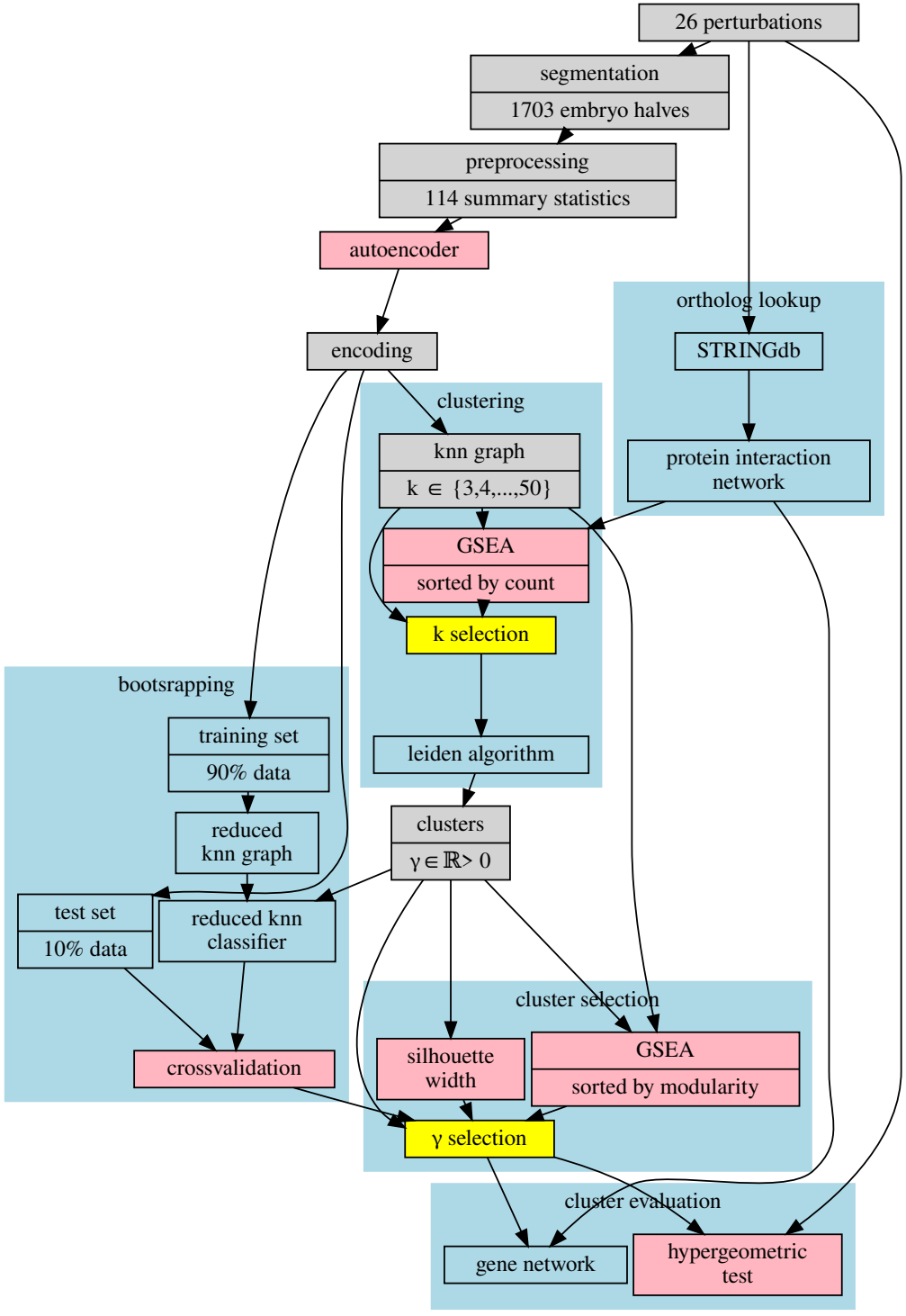
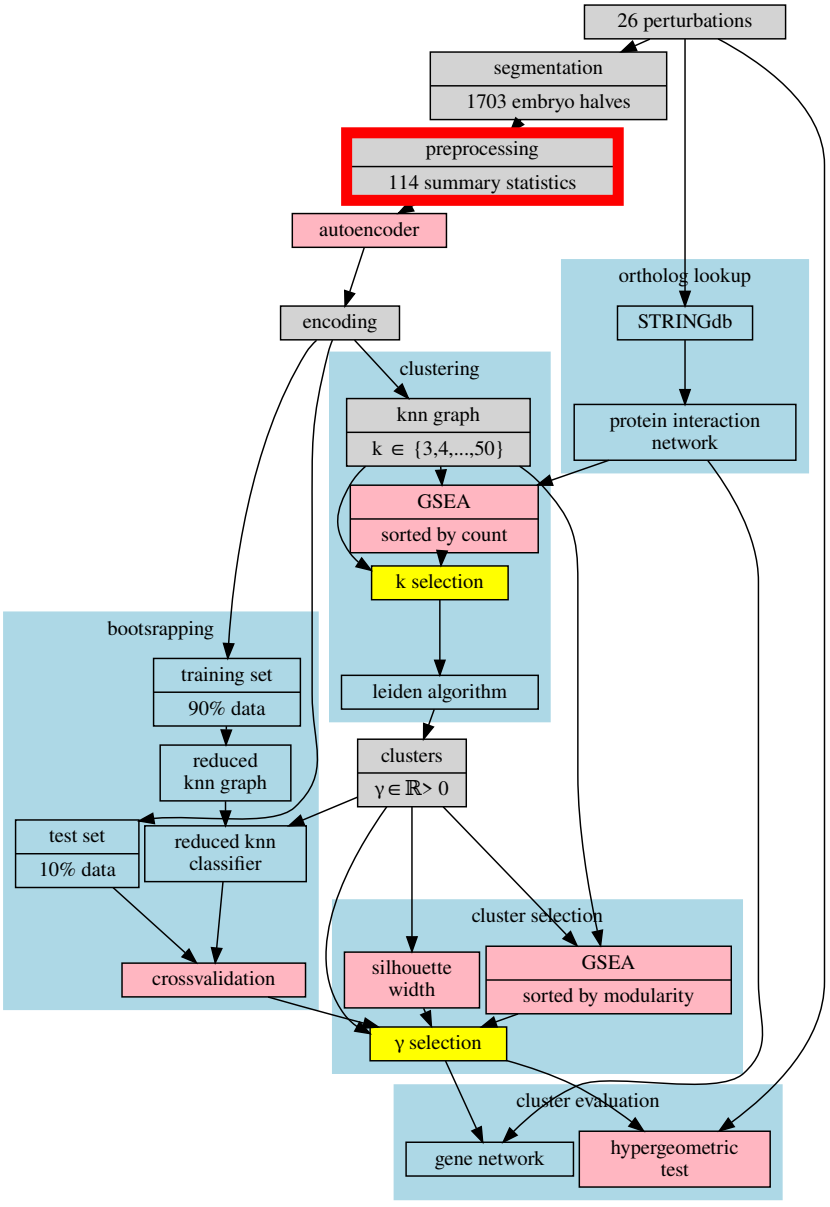
CRISPR Screen

Unsupervised Clustering with Automated Hyperparameter Selection

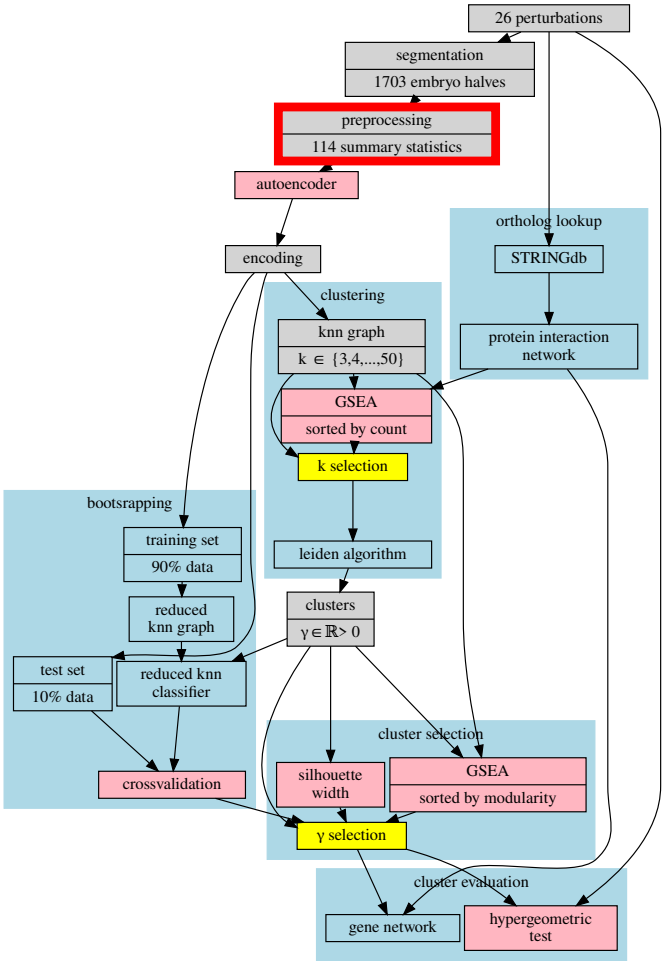
Keira Wiechecki

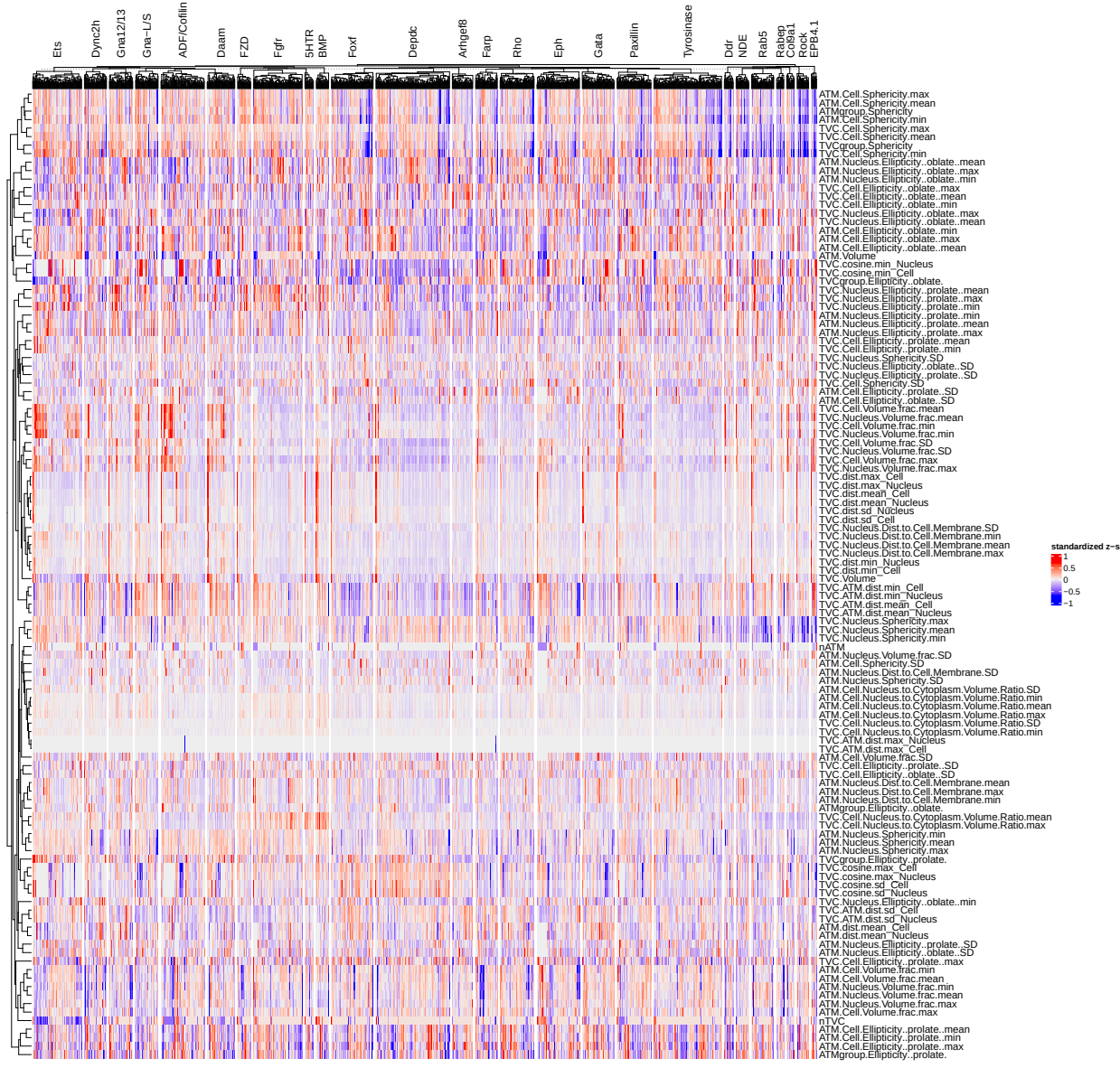
October 23, 2022

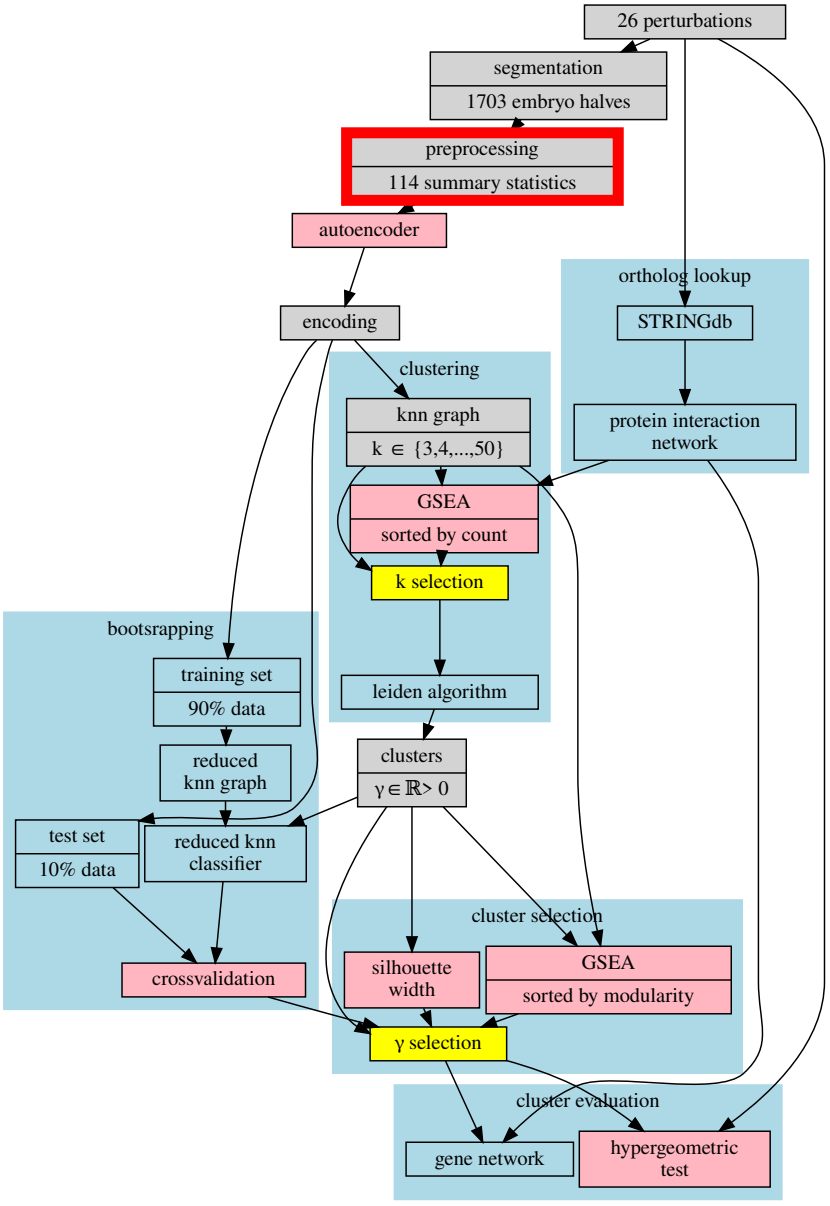

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
bootsrapping
training set
90% data
reduced knn graph
clusters
γ ∈ ℝ> 0
test set
10% data
reduced knn classifier
cluster selection
silhouette width
GSEA
sorted by modularity
crossvalidation
γ selection
cluster evaluation
gene network
hypergeometric test



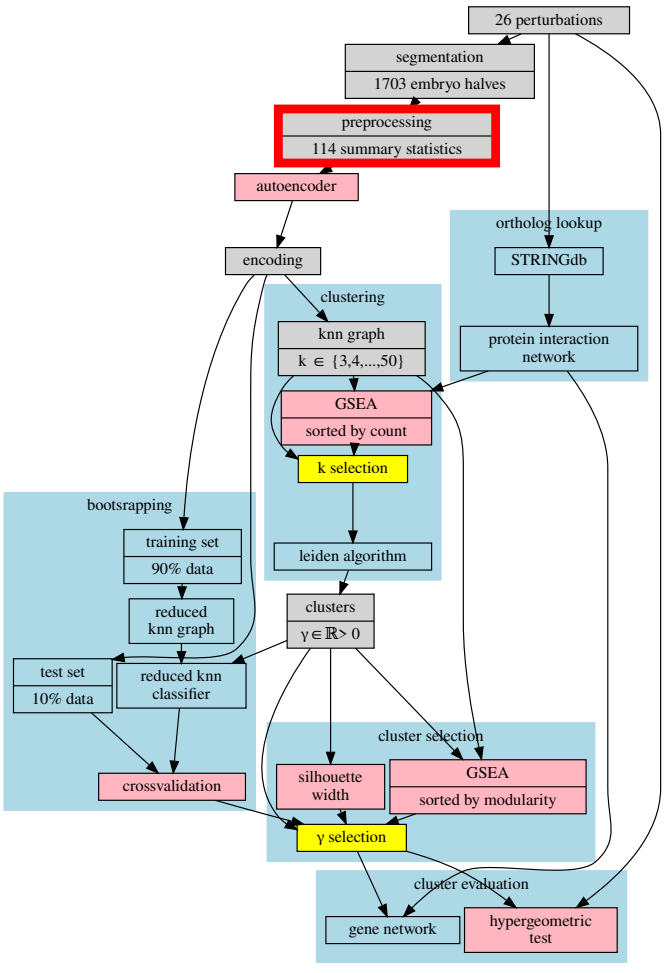
From the segmentation summary statistics I computed 114 embryo-level parameters. Parameters were normalized by z-score then scaled between -1 and 1.

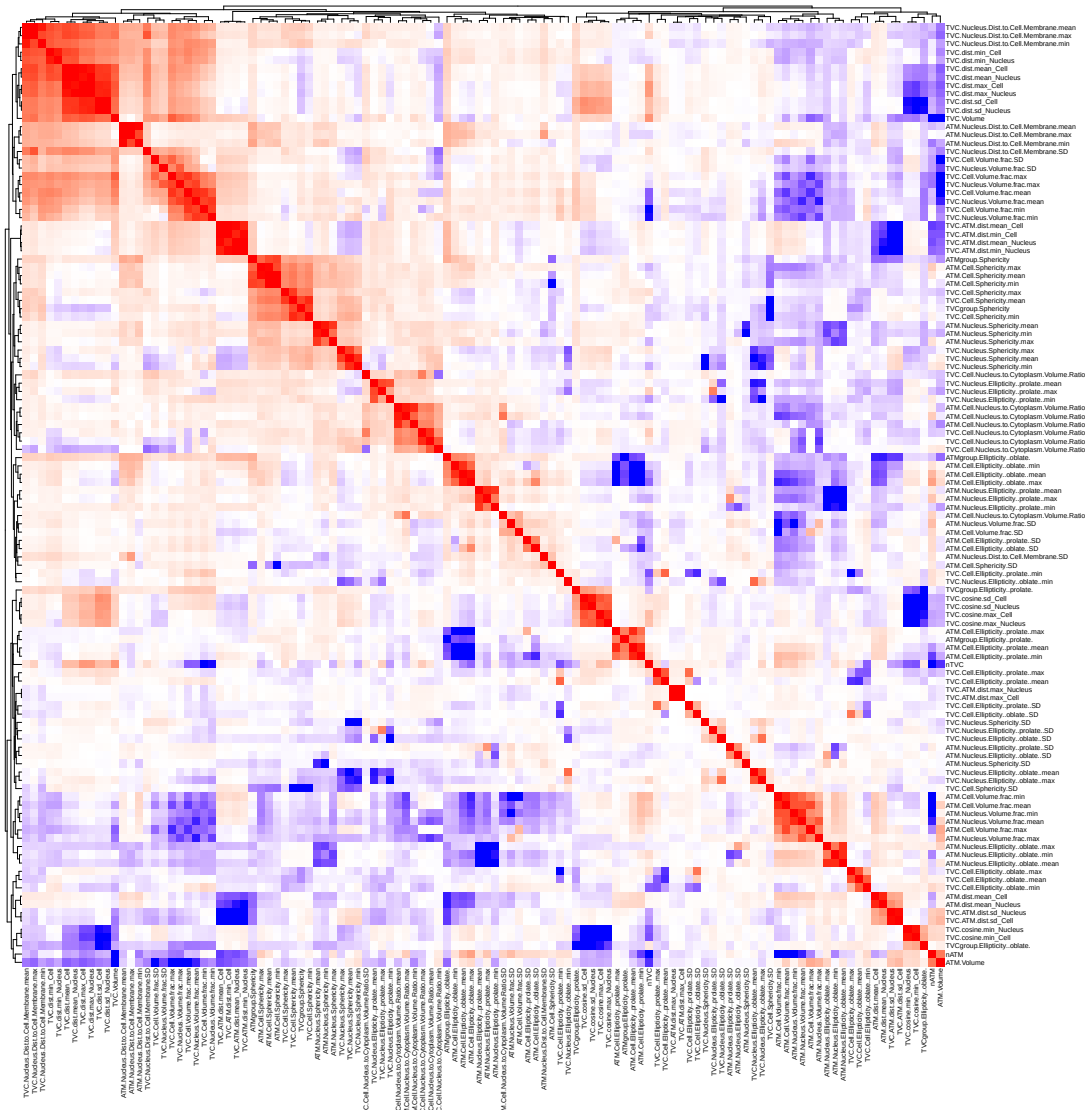

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

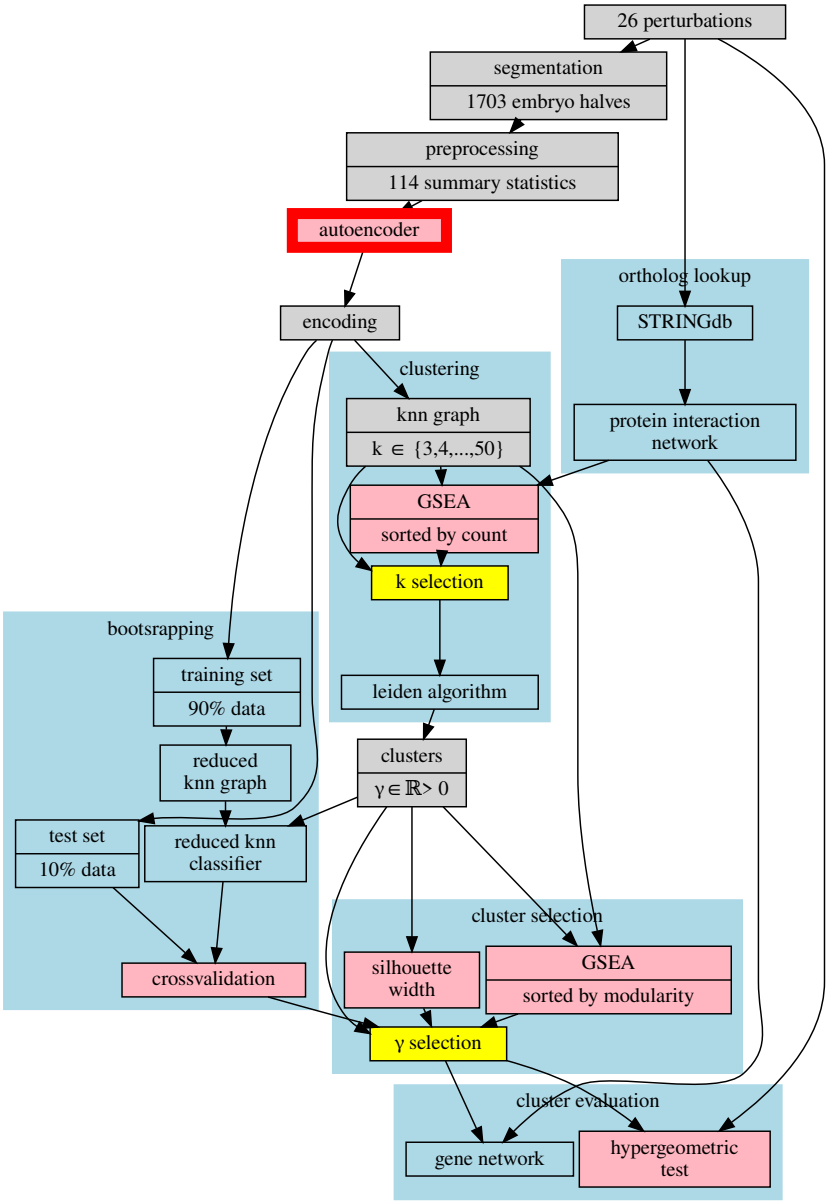

Ets
Dync2h
Gna12/13
Gna−L/S
ADF/Cofilin
Daam
FZD
Fgfr
5HTR BMP
Foxf
Depdc
Arhgef8
Farp
Rho
Eph
Gata
Paxillin
Tyrosinase
Ddr
NDE
Rab5
Rabgp
Cdc42
Rock
EPB4.1
standardized z−s
1
0.5
0
−0.5
−1



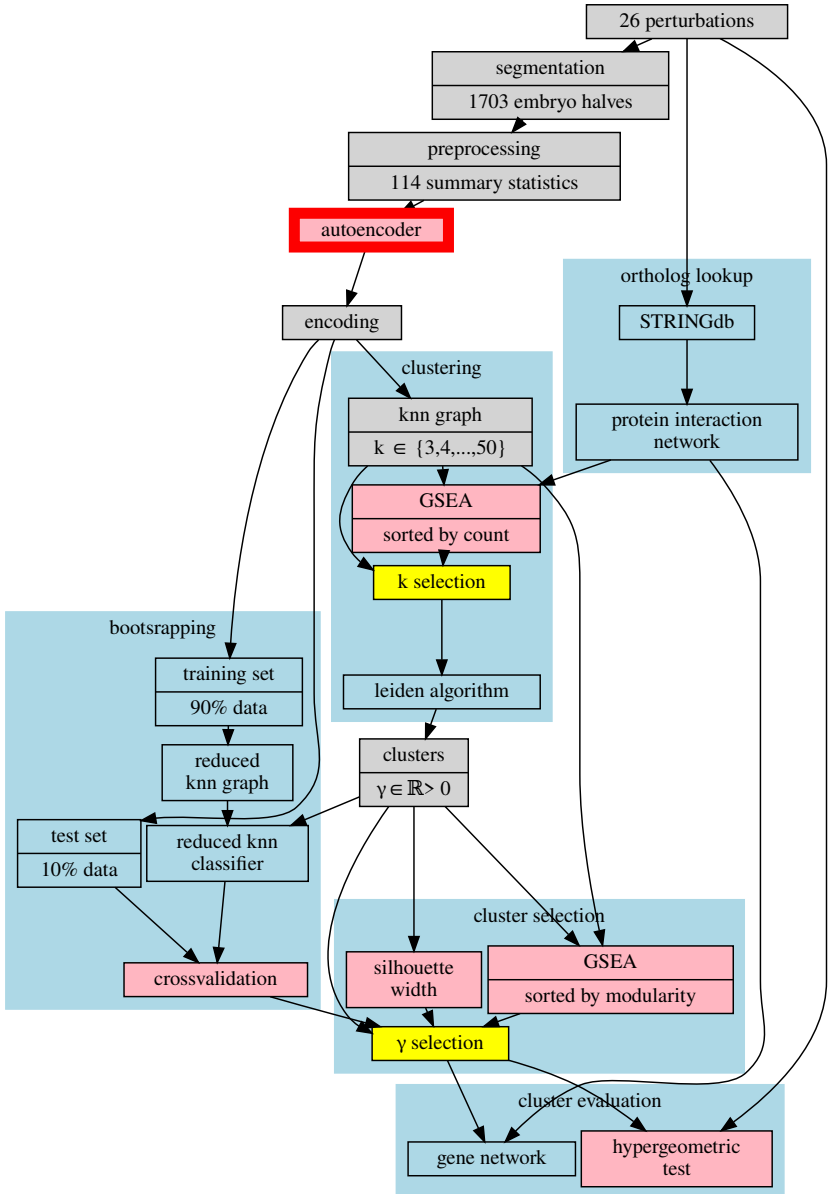
Many parameters are strongly correlated. This is undesirable because each parameter additively contributes to distance used for clustering, resulting in disproportionate weight being given to phenotypes captured by multiple parameters.

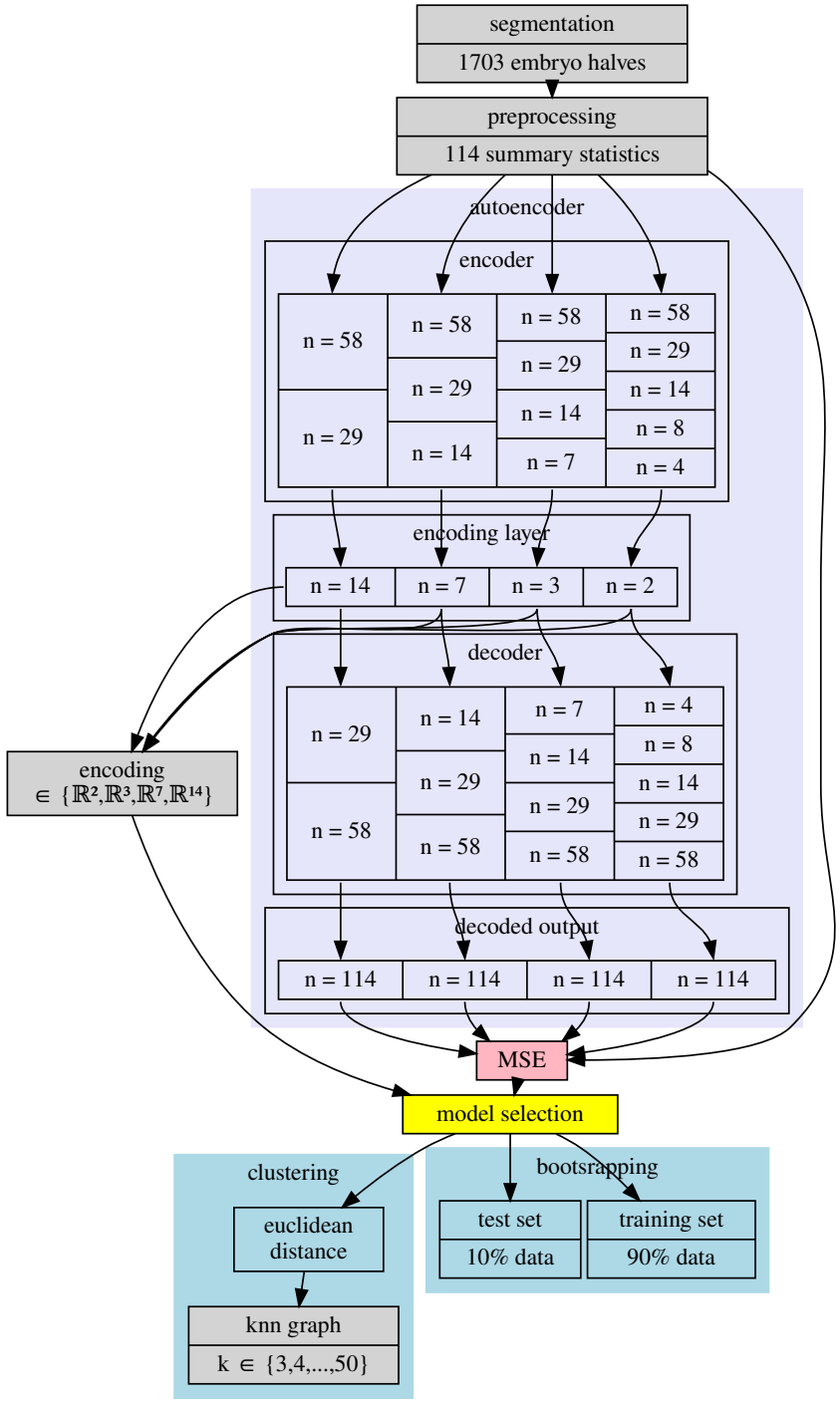

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

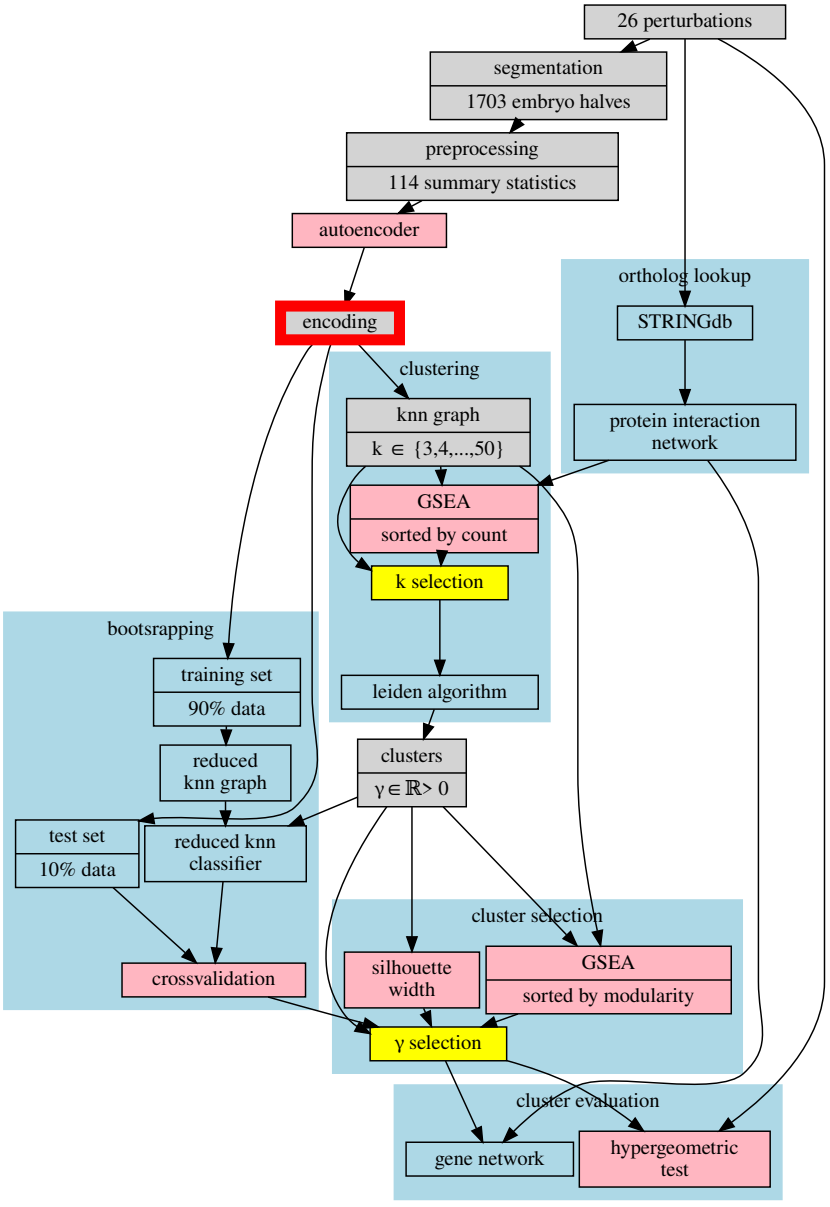




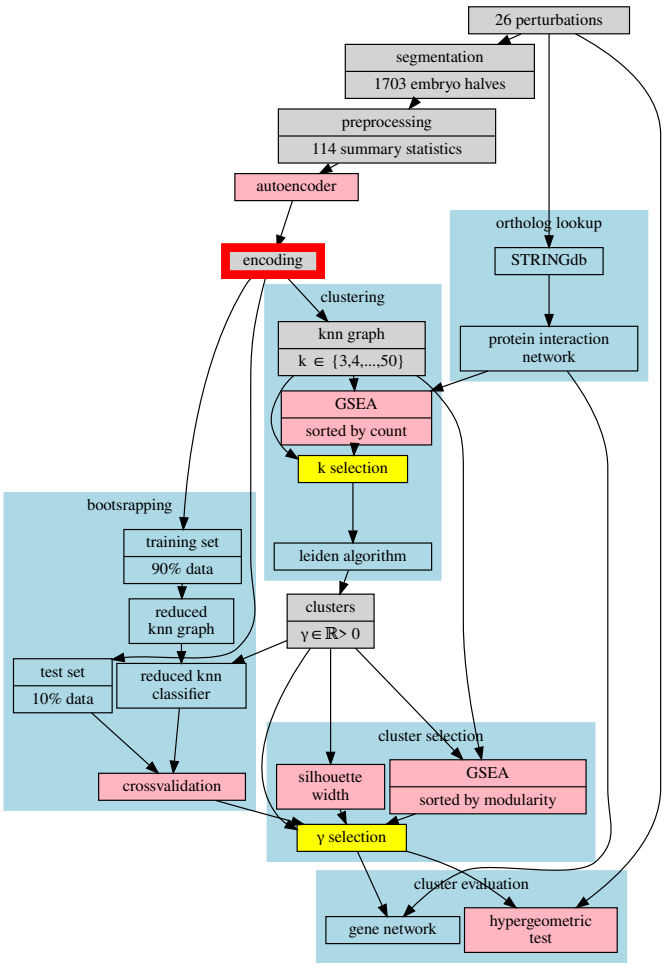
An autoencoder is a method of dimension reduction that uses a neural network to find a lower dimensional encoding which can be decoded to recover the input. This reduces exaggeration of distance due to the number of parameters measured.

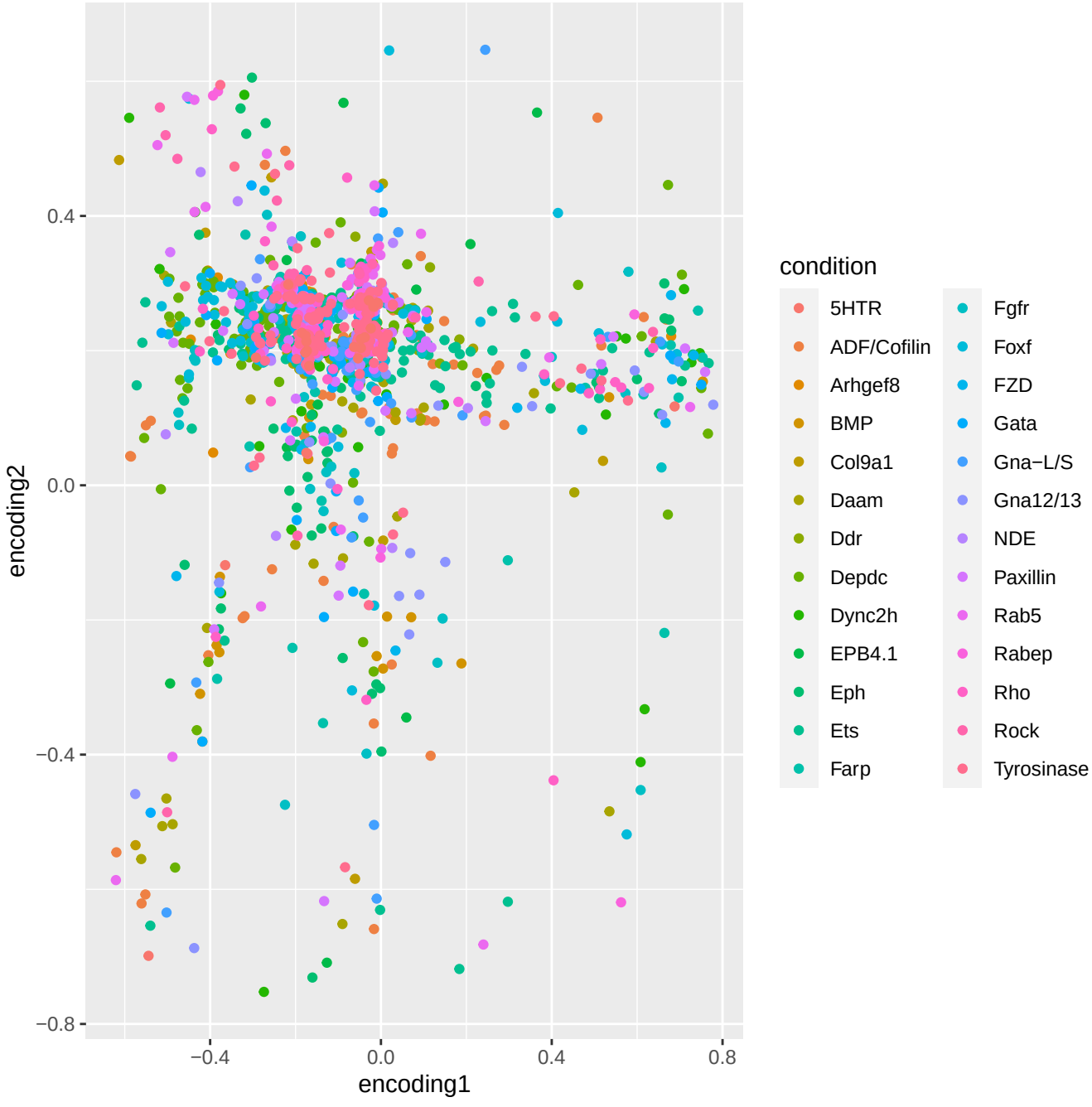

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

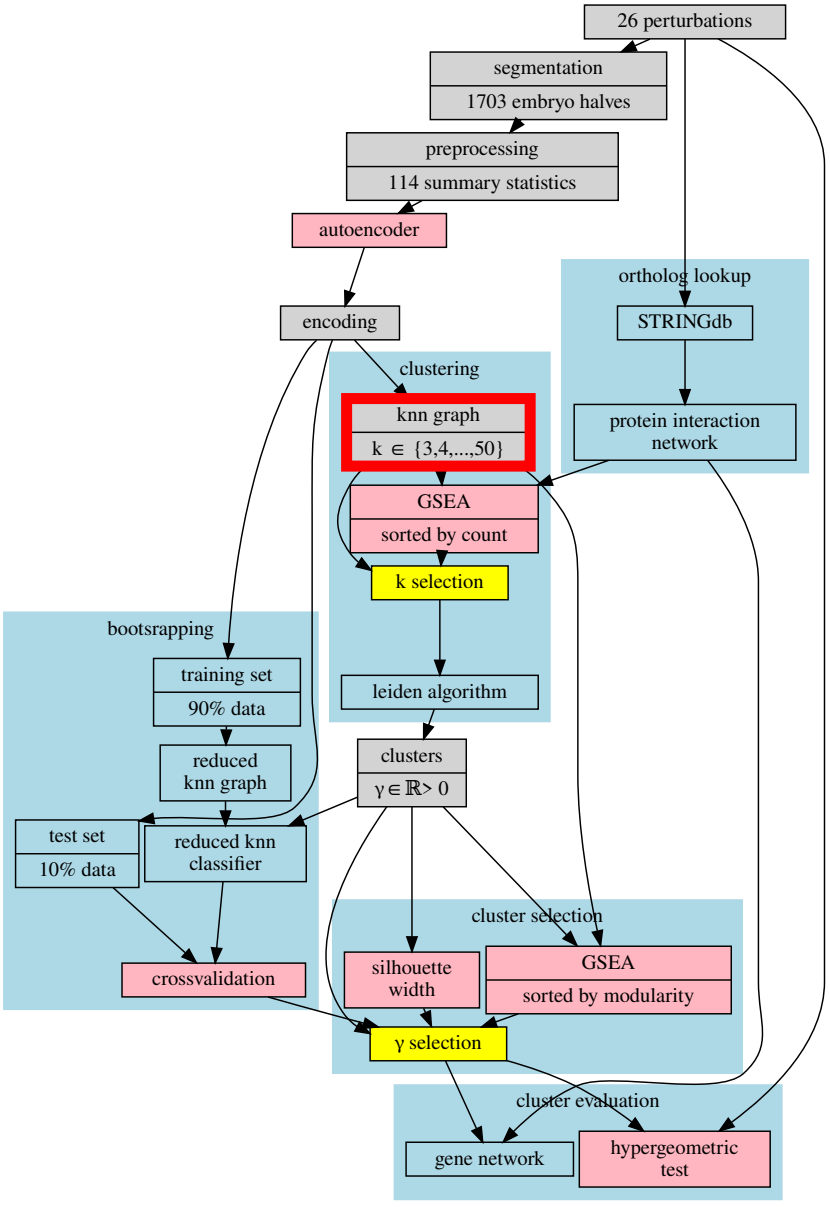

segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoder
n = 58
n = 58
n = 58
n = 58
n = 29
n = 29
n = 29
n = 29
n = 14
n = 14
n = 14
n = 8
n = 7
n = 4
encoding layer
n = 14
n = 7
n = 3
n = 2
decoder
n = 29
n = 14
n = 7
n = 4
n = 8
n = 14
n = 29
n = 14
n = 29
n = 29
n = 58
n = 58
n = 58
n = 58
encoding
∈ {ℝ²,ℝ³,ℝ⁷,ℝ¹⁴}
decoded output
n = 114
n = 114
n = 114
n = 114
MSE
model selection
clustering
euclidean distance
knn graph
k ∈ {3,4,...,50}
bootsrapping
test set
10% data
training set
90% data



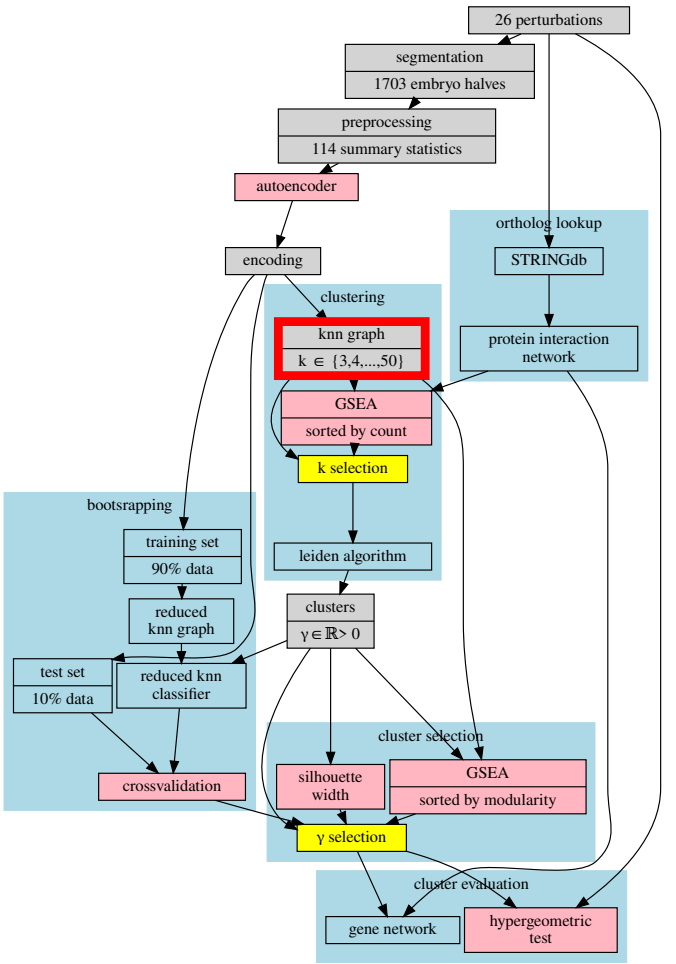
More than 2D produced only marginal improvement.

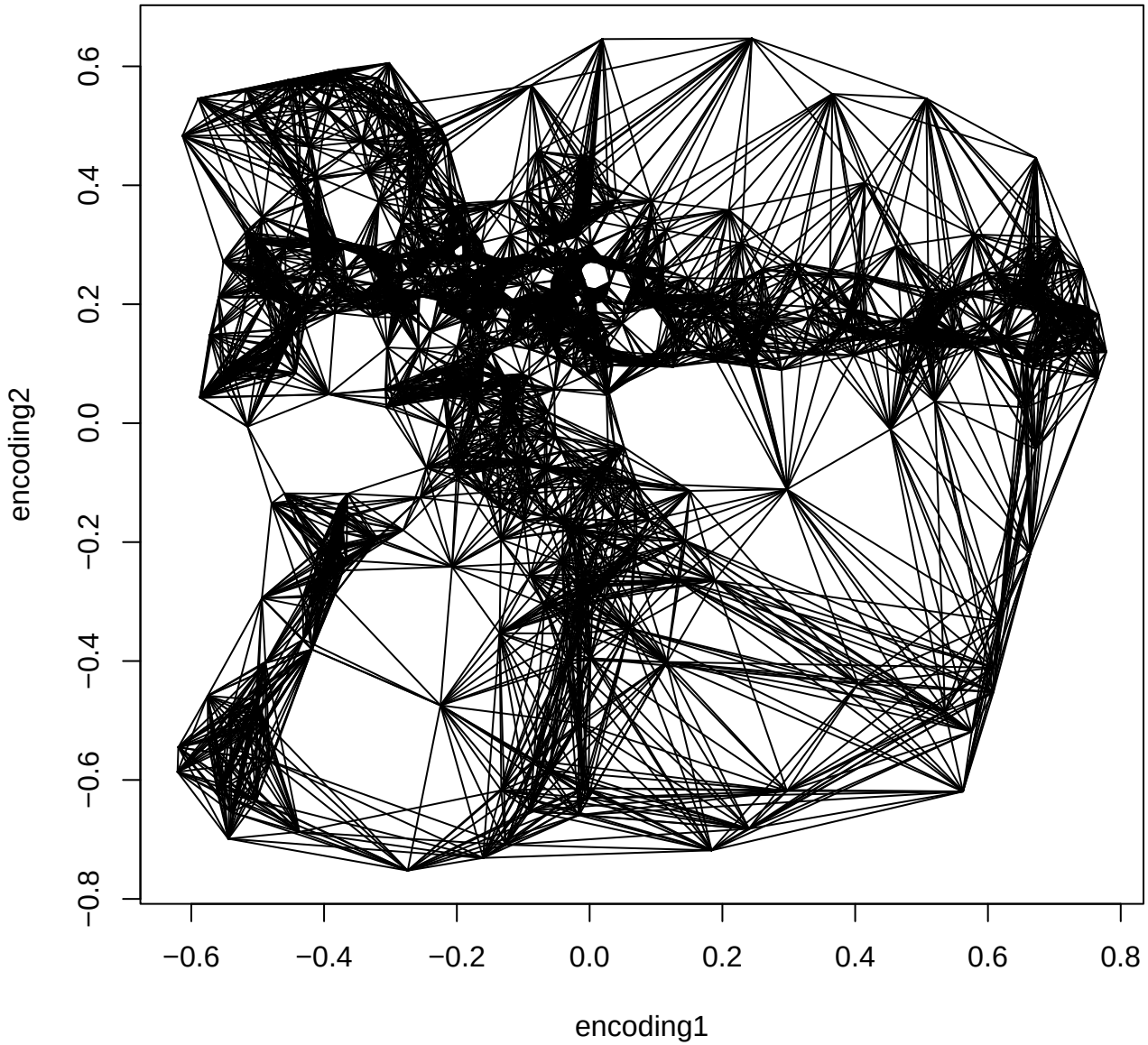

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

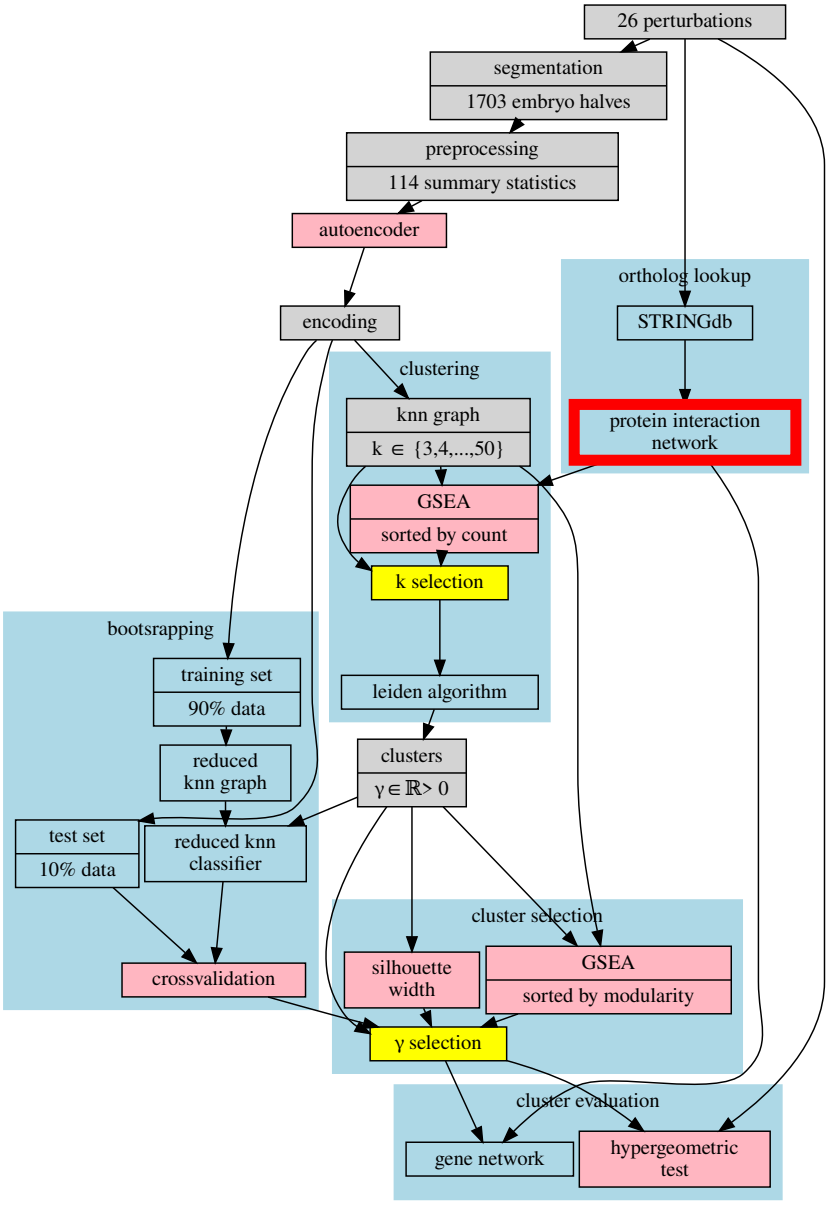

encoding2
encoding1
0.4
0.0
−0.4
−0.8
−0.4
0.0
0.4
0.8
condition
5HTR
ADF/Cofilin
Arhgef8
BMP
Col9a1
Daam
Ddr
Depdc
Dync2h
EPB4.1
Eph
Ets
Farp
Fgfr
Foxf
FZD
Gata
Gna−L/S
Gna12/13
NDE
Paxillin
Rab5
Rabep
Rho
Rock
Tyrosinase



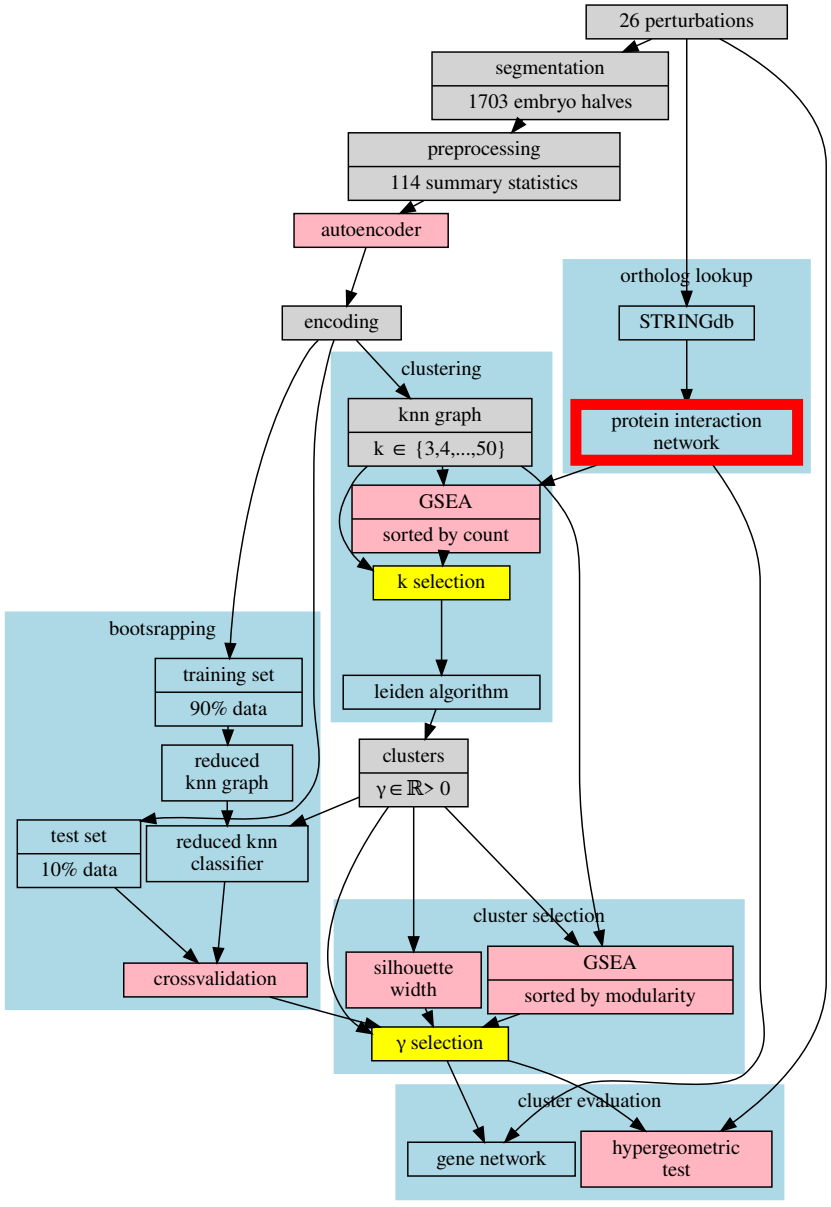
A kNN graph is constructed from euclidean distance calculated from the encoding layer.

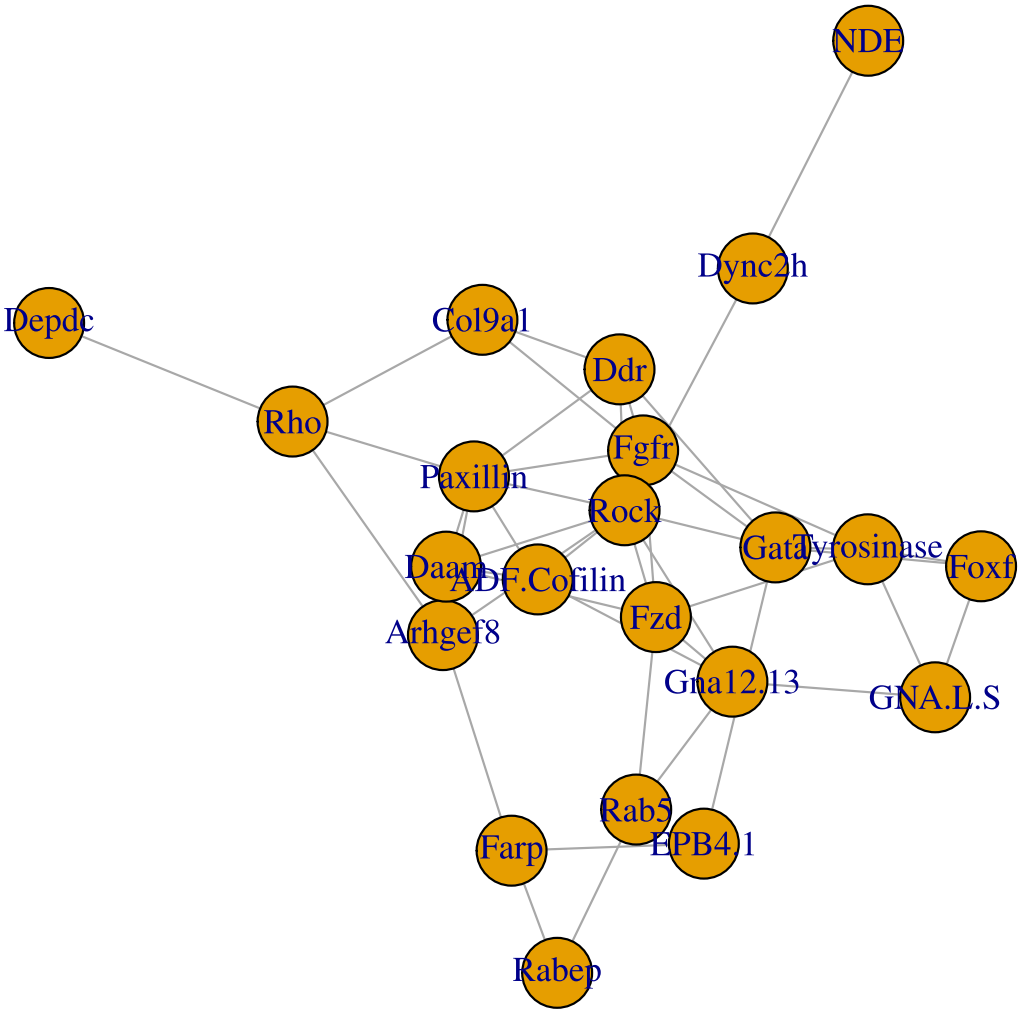

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

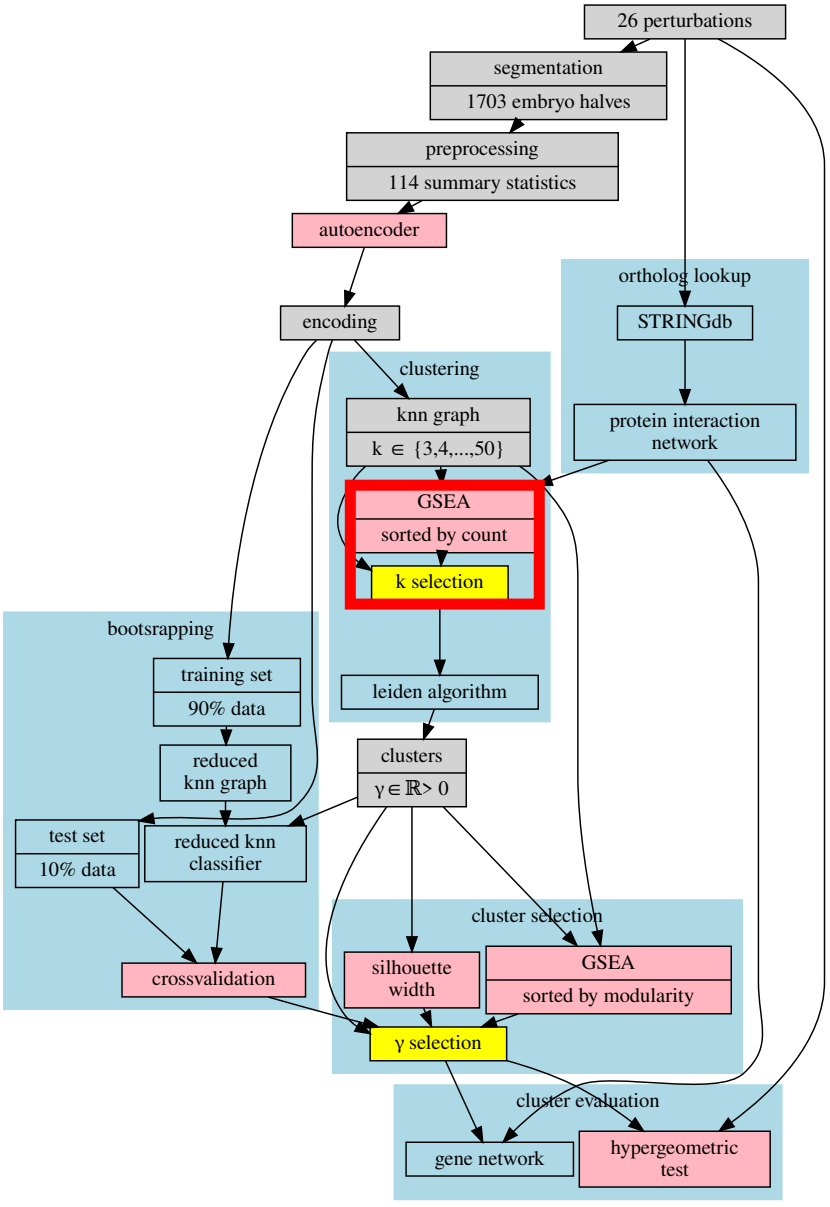

encoding2
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
encoding1



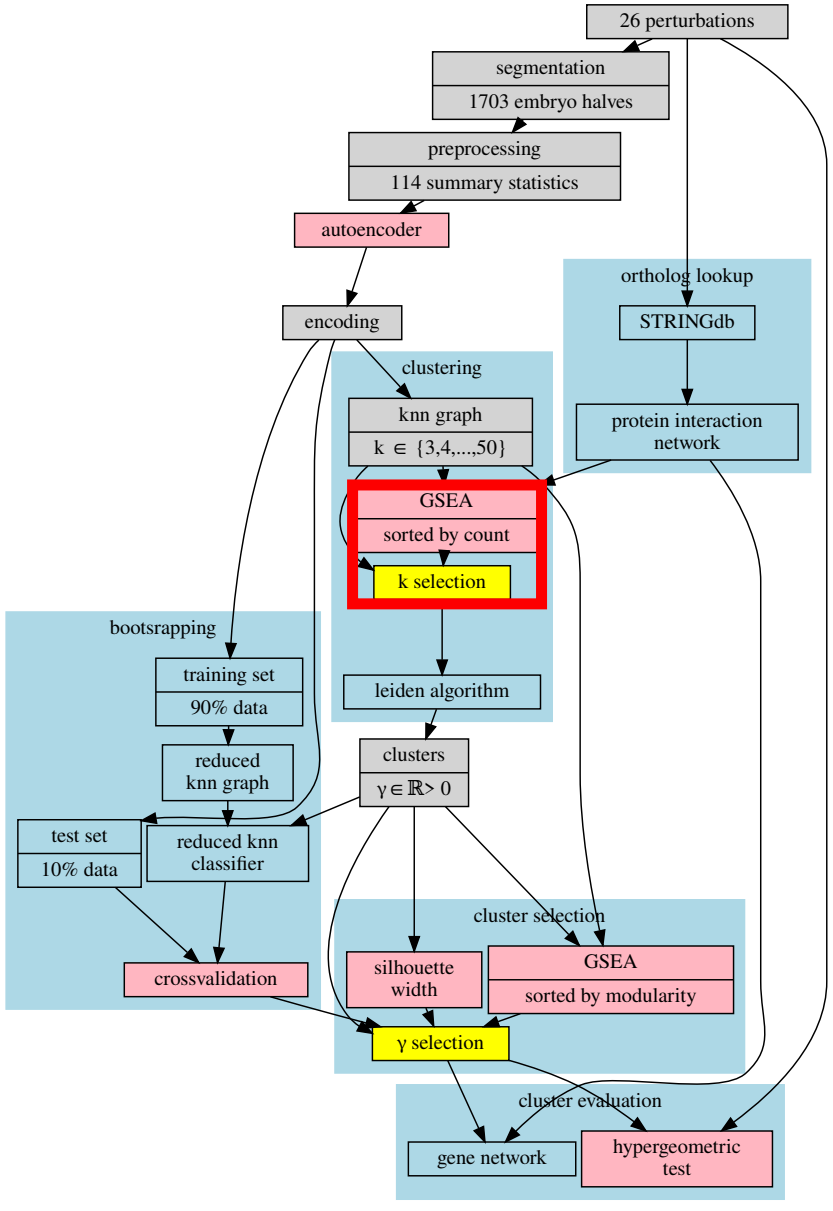
I obtained orthologs from *M. musculus* and *H. sapiens* from ENSEMBL. I used STRINGdb to obtain known protein interactions.

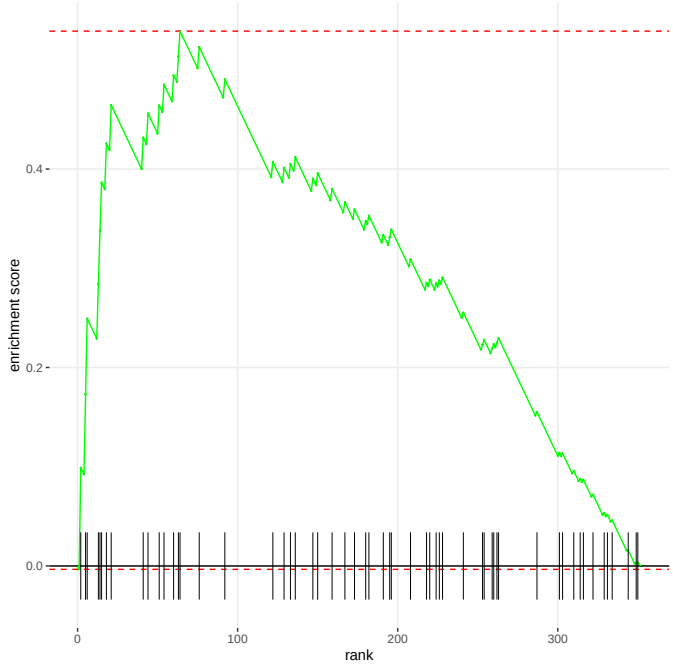

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

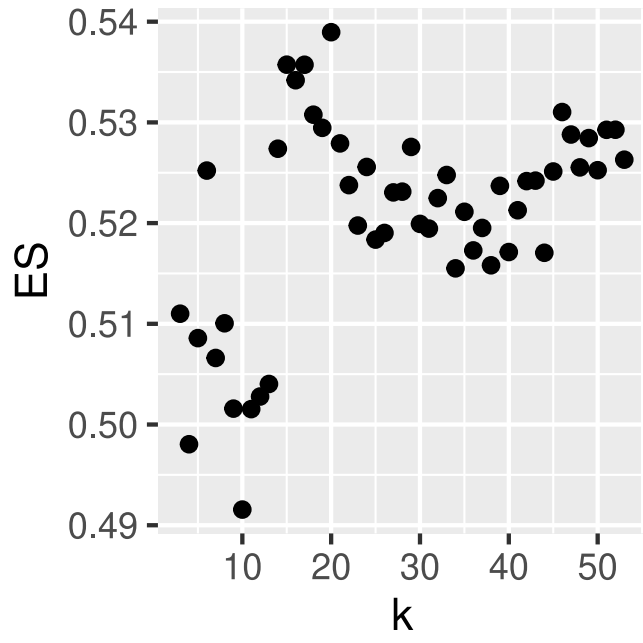

NDE
Dync2h
Depdc
Col9a1
Ddr
Rho
Fgfr
Paxillin
Rock
Tyrosinase
Foxf
Daam
ADF.Cofilin
Fzd
Arhgef8
Gna12.13
GNA.L.S
Rab5
Farp
EPB4.1
Rabep

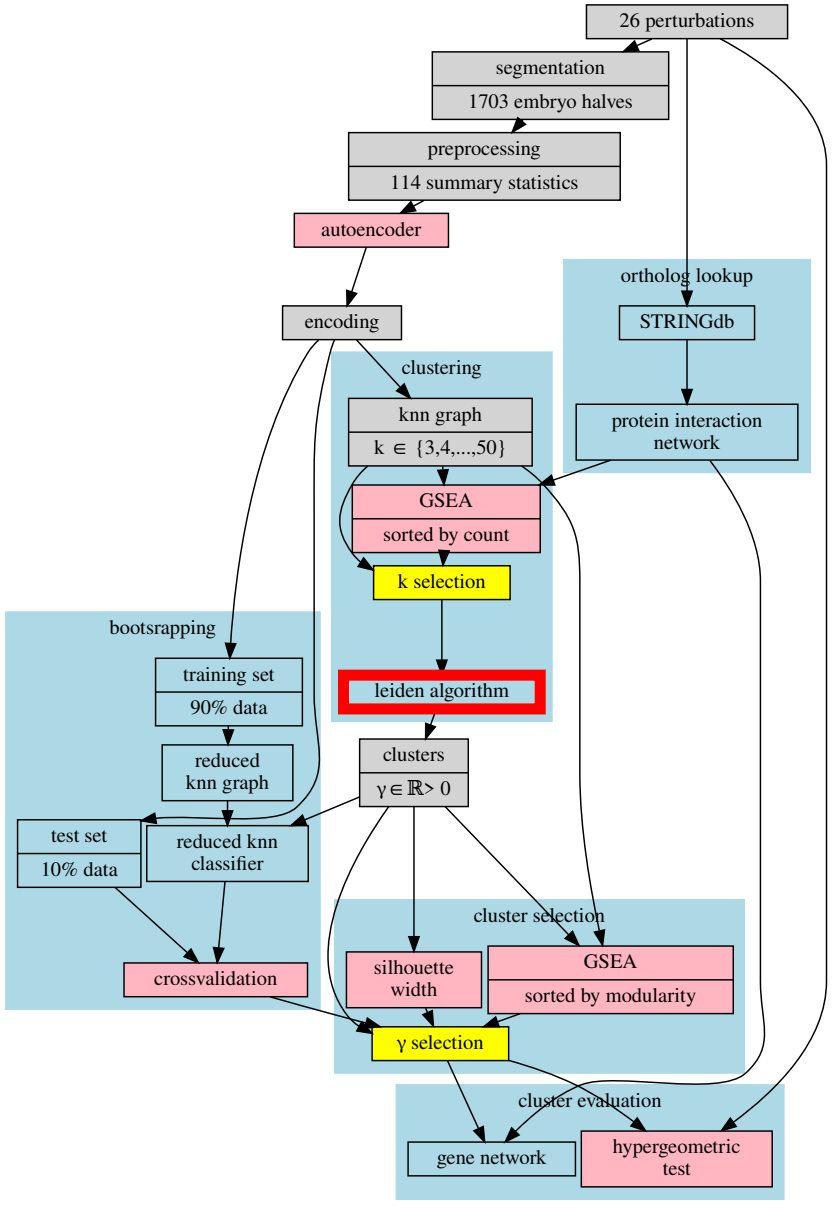


The known protein interactions can be treated as a gene set for GSEA. Interactions can be ranked by edge count between embryos in two conditions. An enrichment score is calculated based on occurrence of known interactions near the top of the ranked list. An optimal k can be selected by maximizing enrichment score.


26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
clusters
γ ∈ ℝ> 0
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test


enrichment score
0.4
0.2
0.0
0
100
200
300
rank

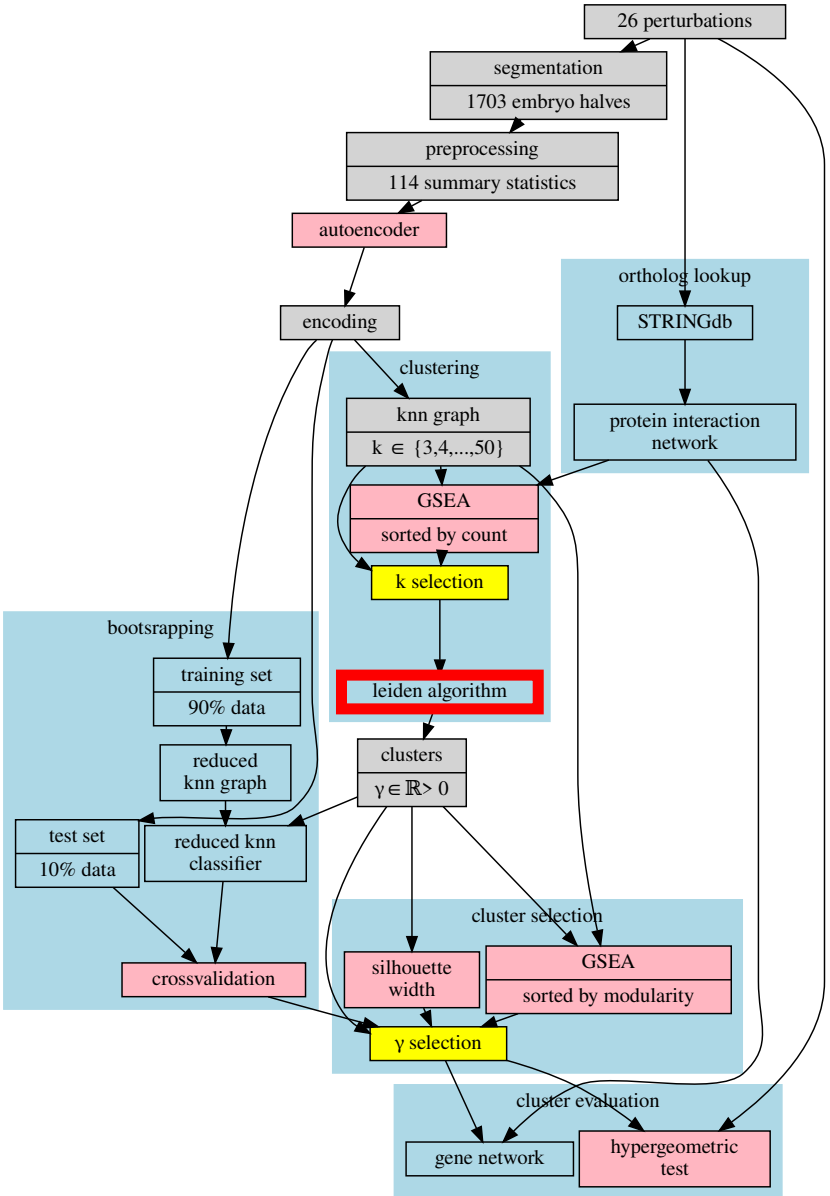

ES
0.54
0.53
0.52
0.51
0.50
0.49
10
20
30
40
50
k

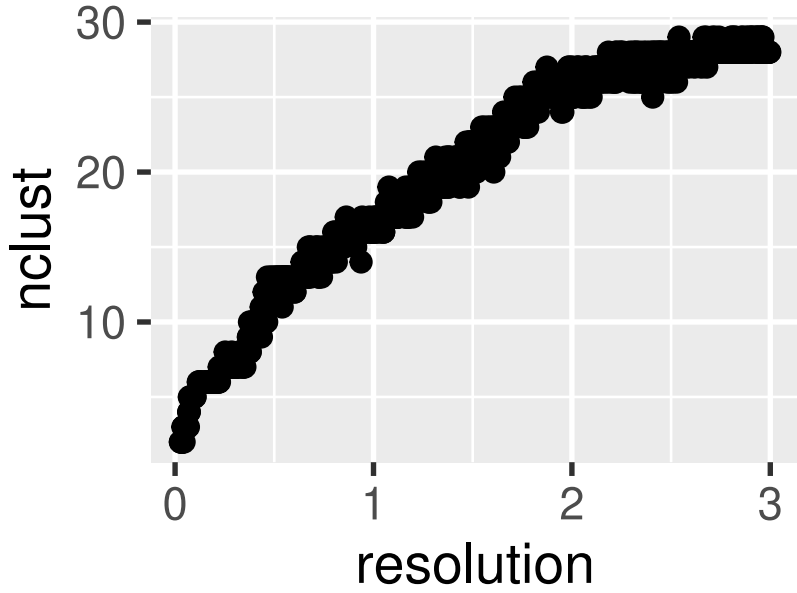


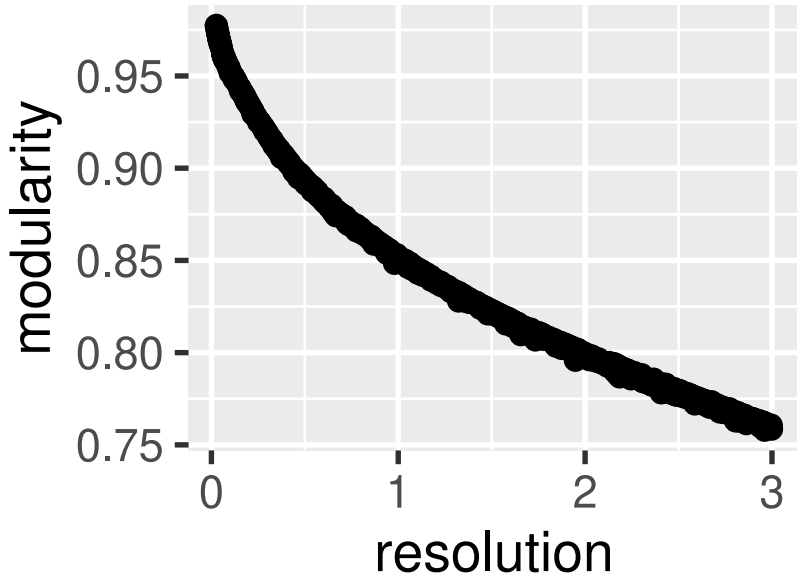
The leiden algorithm attempts to find a clustering that maximizes modularity H for a given graph and resolution γ. Modularity is defined as

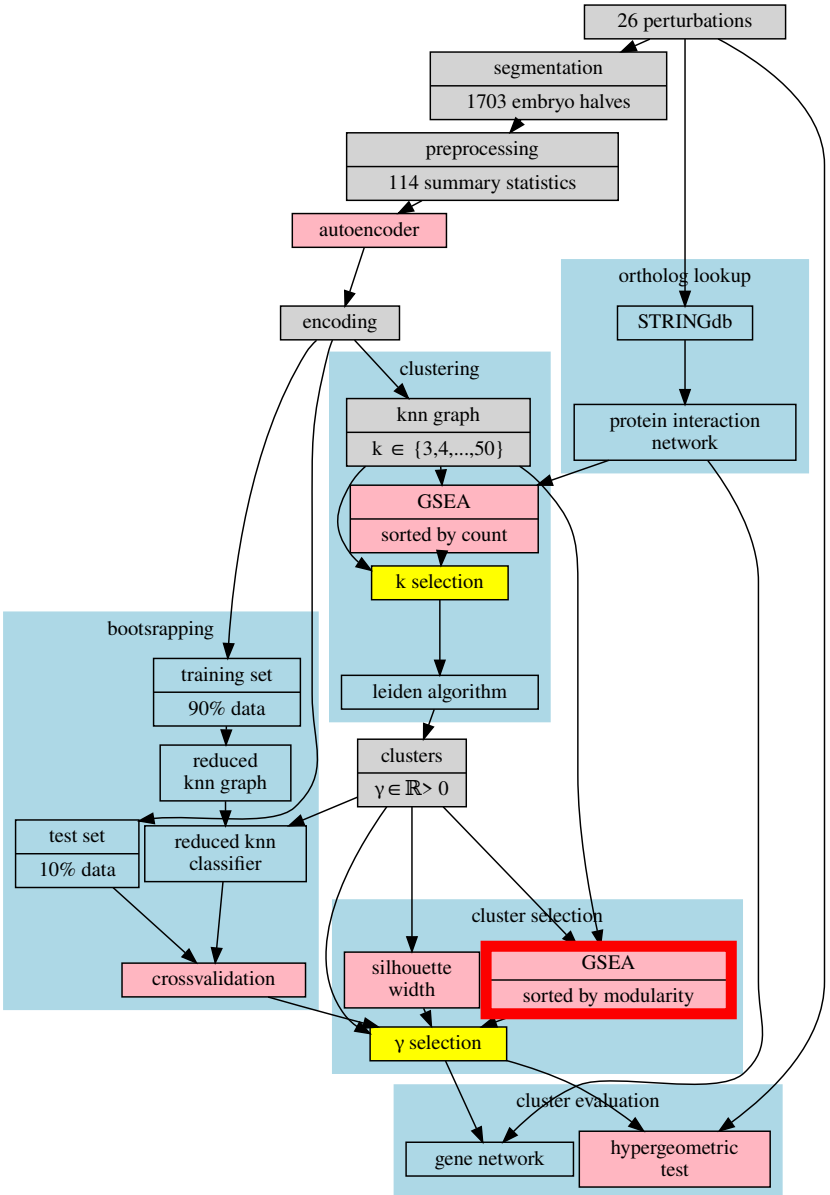
$$H = \frac{1}{2m} \sum_c \left(e_c - \gamma \frac{K_c^2}{2m} \right)$$

where m is the average degree of the graph, e_c is the number of edges in cluster c, and K_c is the number of nodes in cluster c. This gives a measure of how well-connected clusters compared to expectation based on average degree of the graph and number of nodes in a cluster. A higher γ results in more clusters.

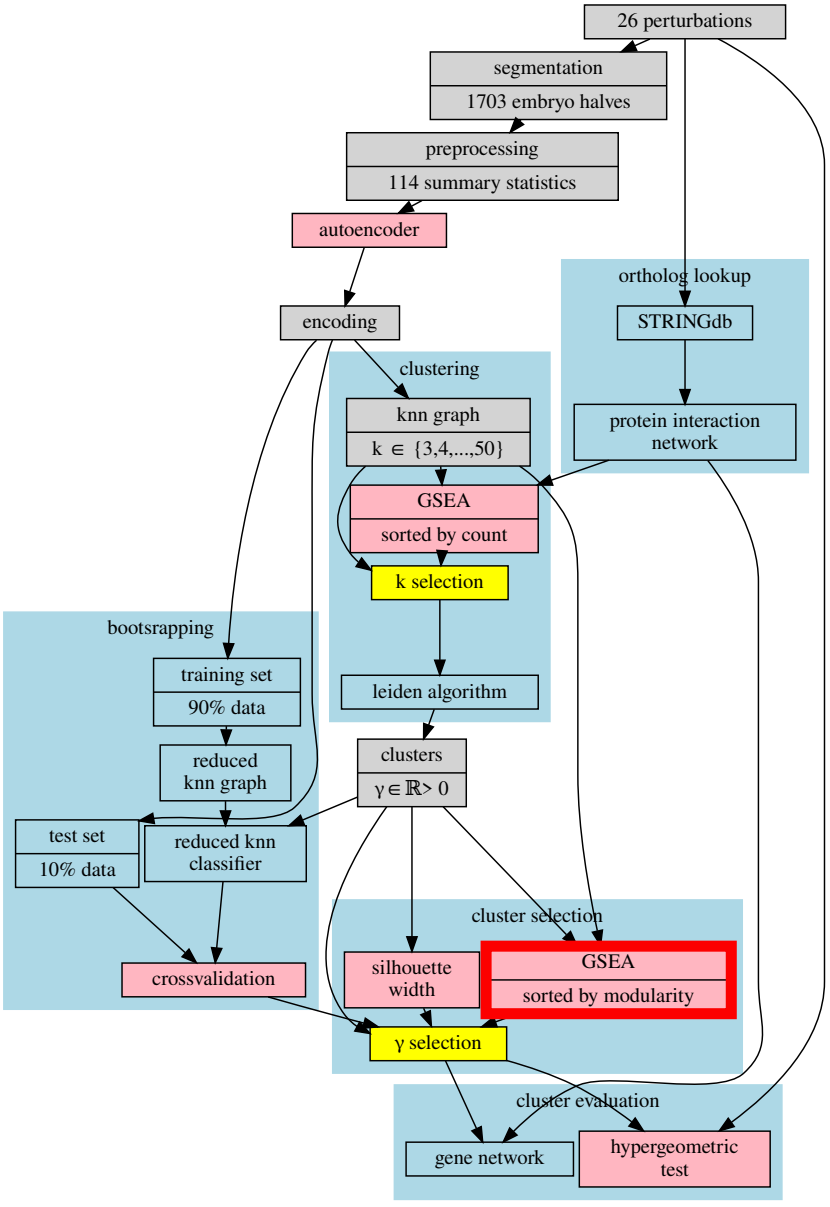

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

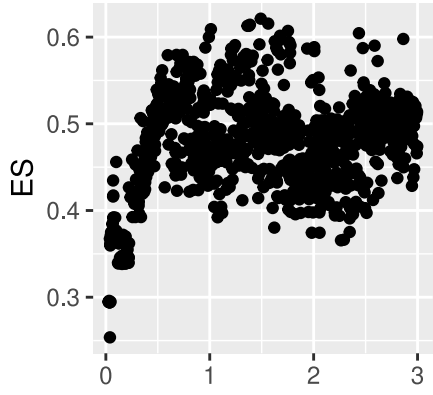

nclust
30
20
10
0
1
2
3
resolution

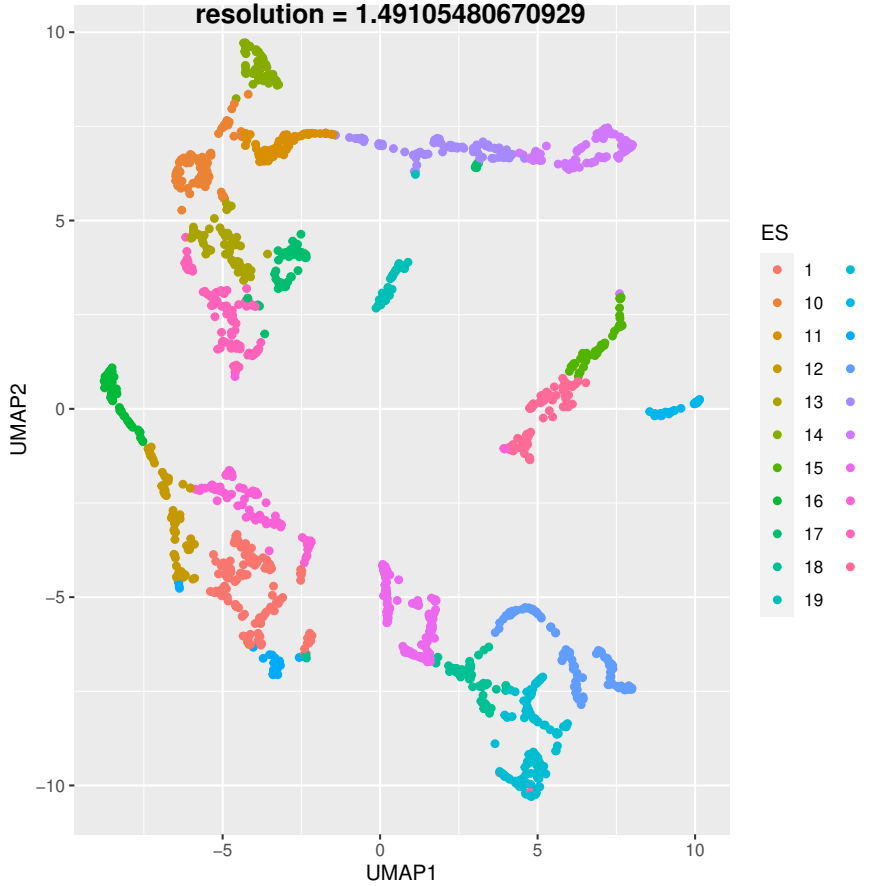

modularity
0.95
0.90
0.85
0.80
0.75
0
1
2
3
resolution

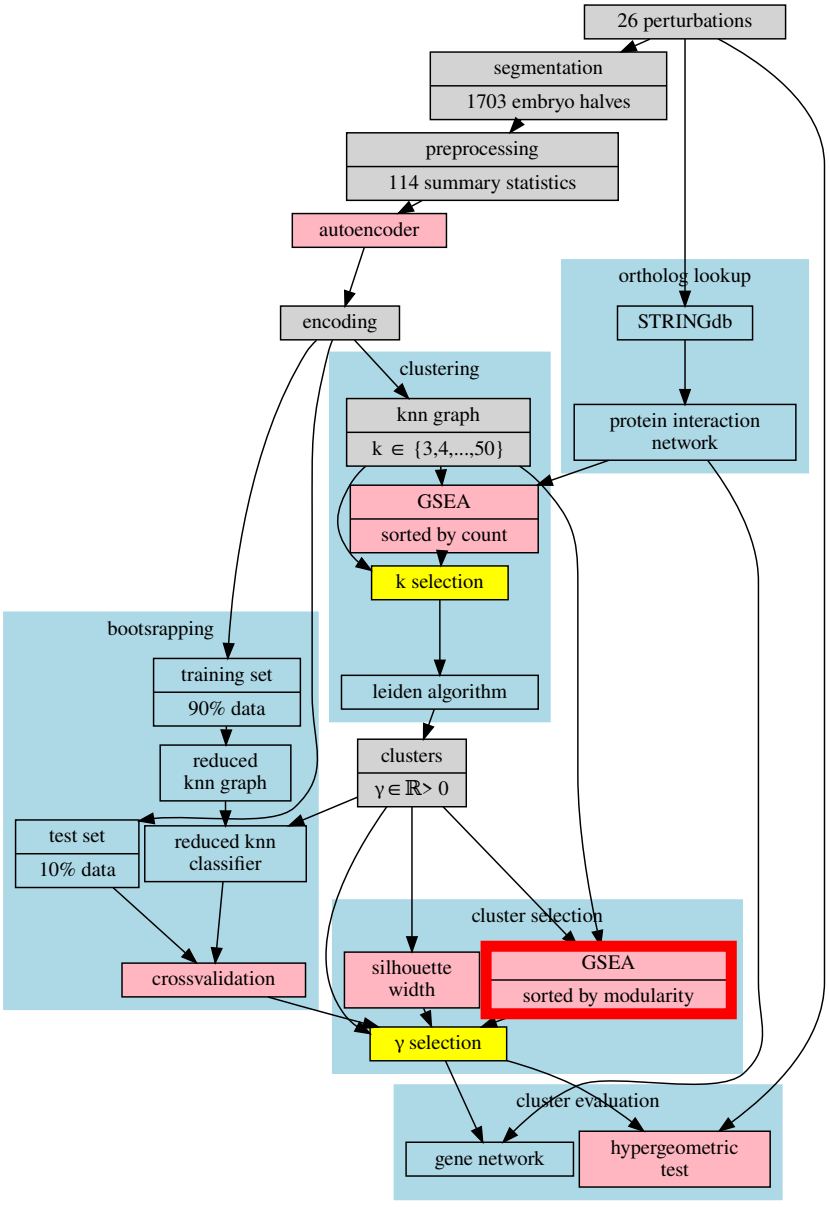


I obtained 1000 clusterings on random γ values between 0.01 and 3. I assessed clustering by several metrics. I obtained an enrichment score from a second GSEA with interactions ranked by modularity of a subgraph consisting of only edges between two given conditions.

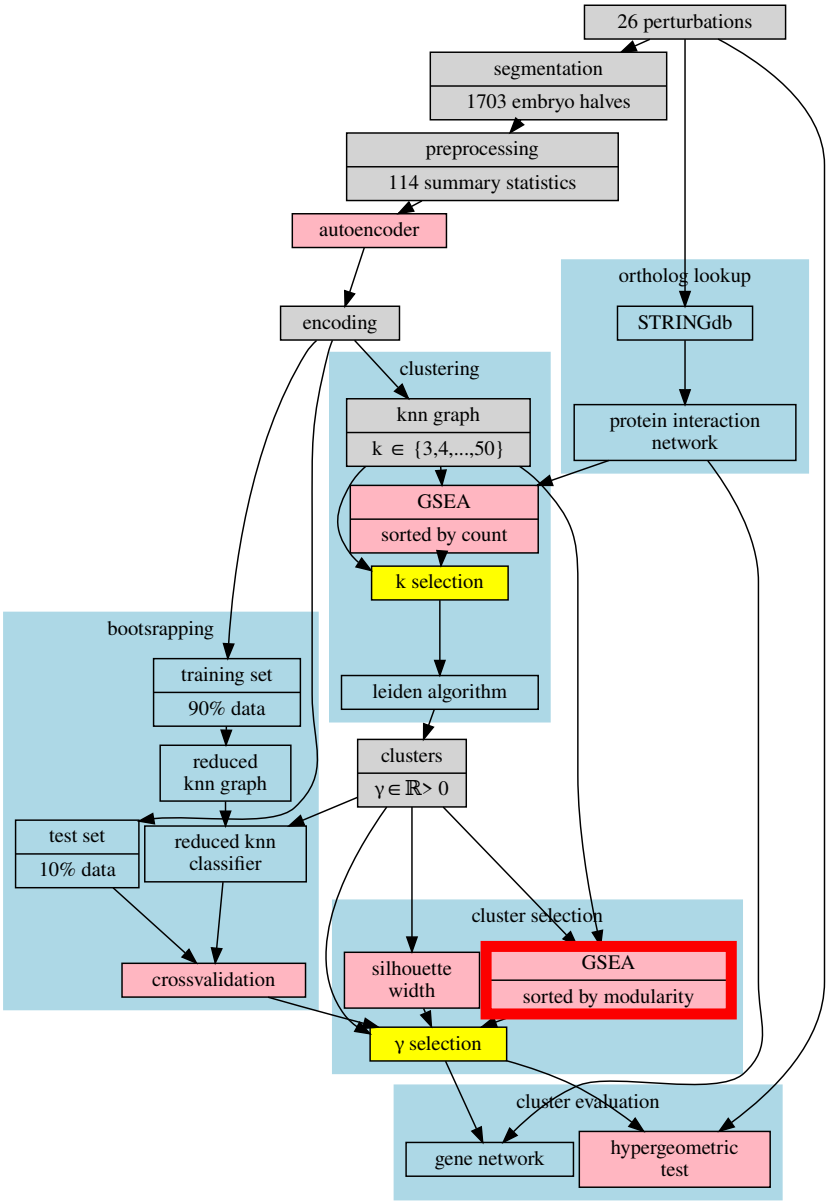

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

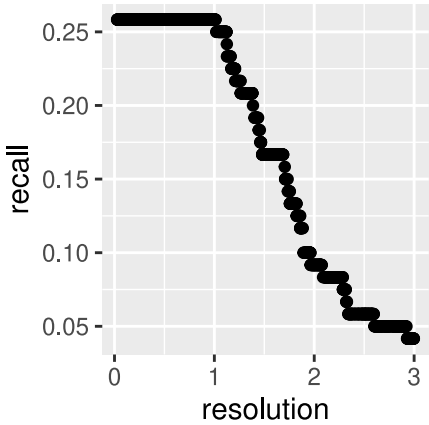

ES
0.6
0.5
0.4
0.3
0
1
2
3
resolution

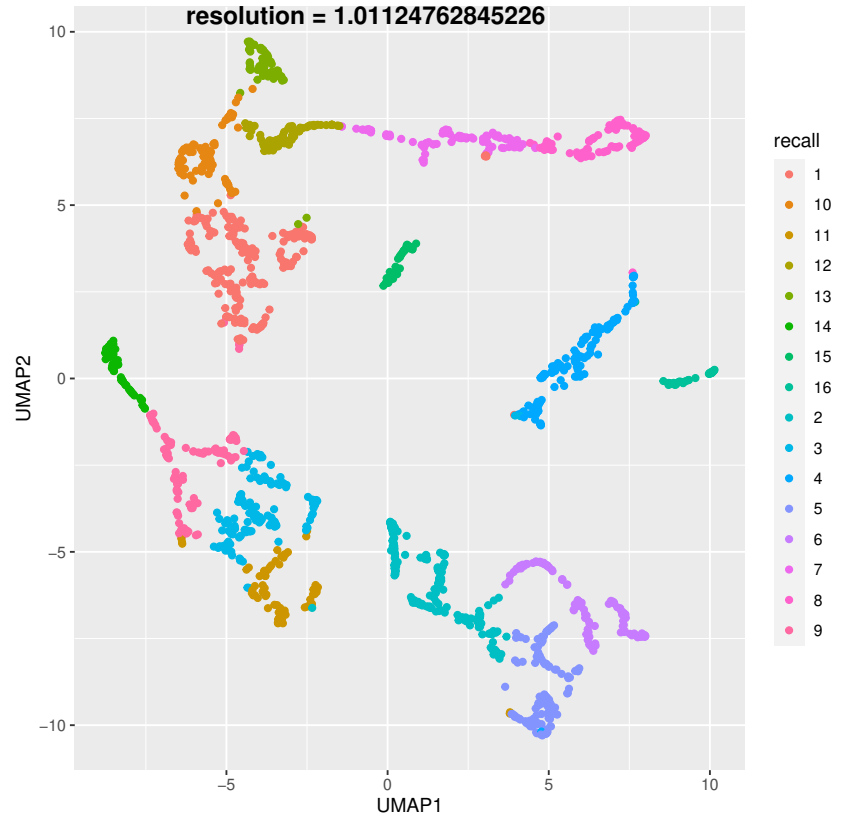

resolution = 1.49105480670929
UMAP2
UMAP1
10
5
0
−5
−10
ES
1
10
11
12
13
14
15
16
17
18
19

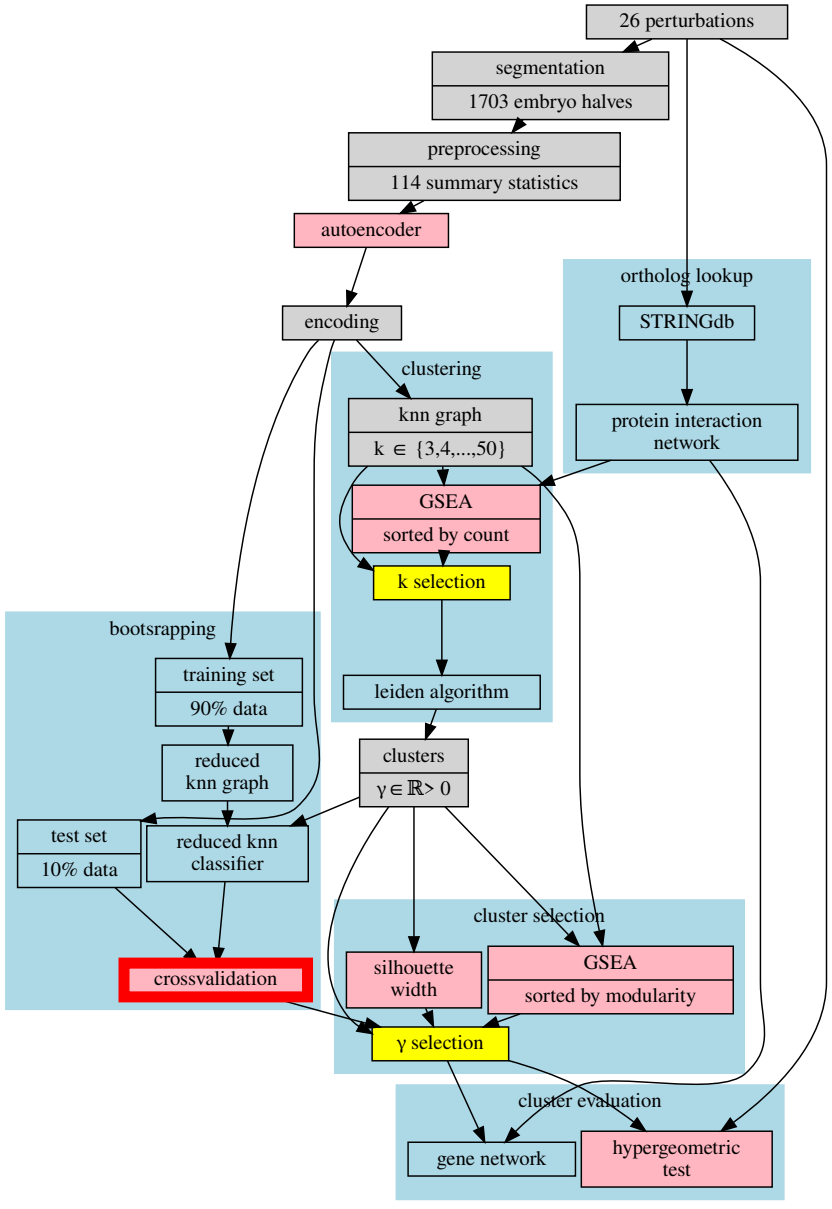


I used this same metric to construct a gene network. If an interaction has positive modularity the genes are considered connected. A recall score can be calculated by comparing this network to the protein interaction network.

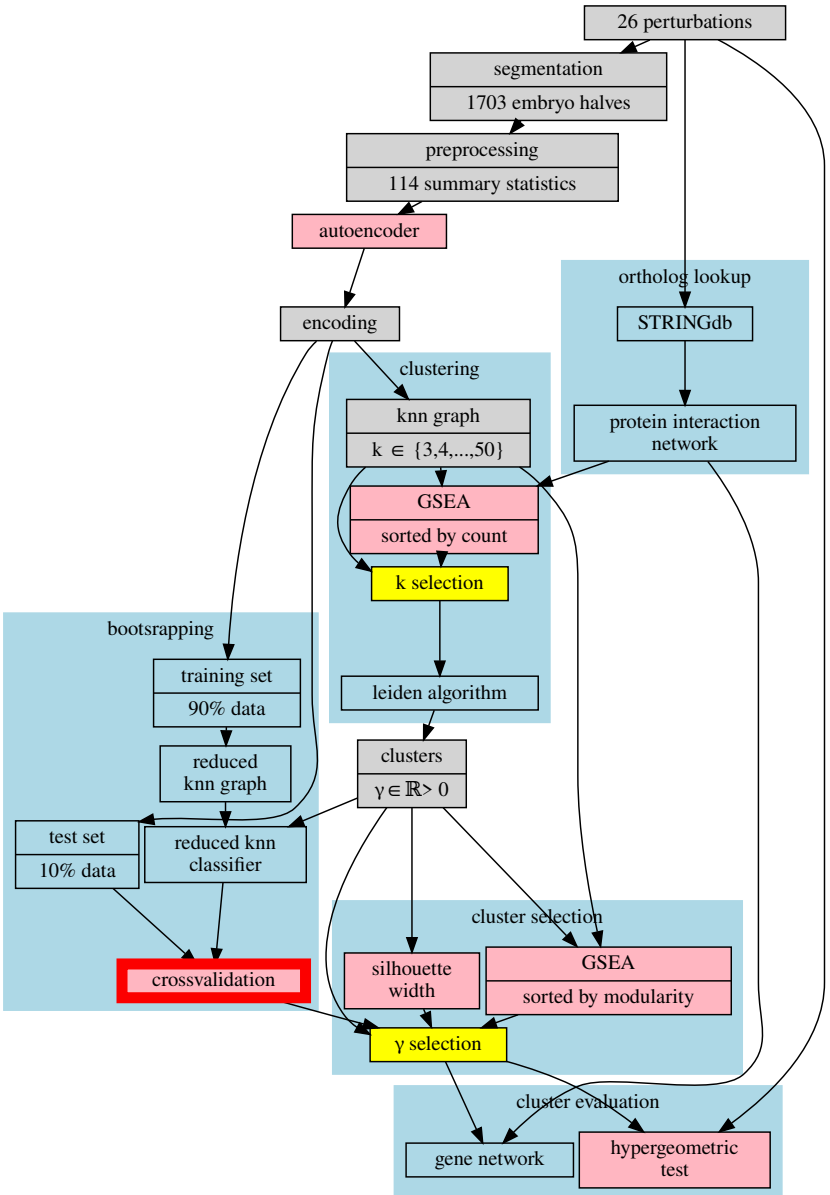

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

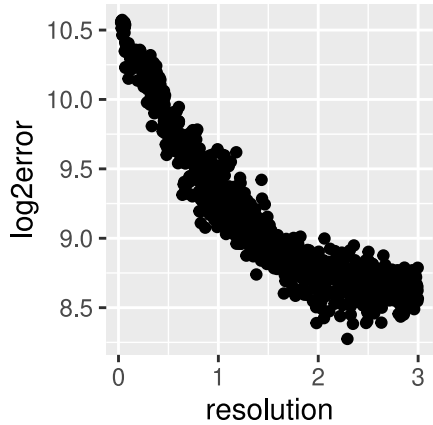

recall
0.25
0.20
0.15
0.10
0.05
0
1
2
3
resolution

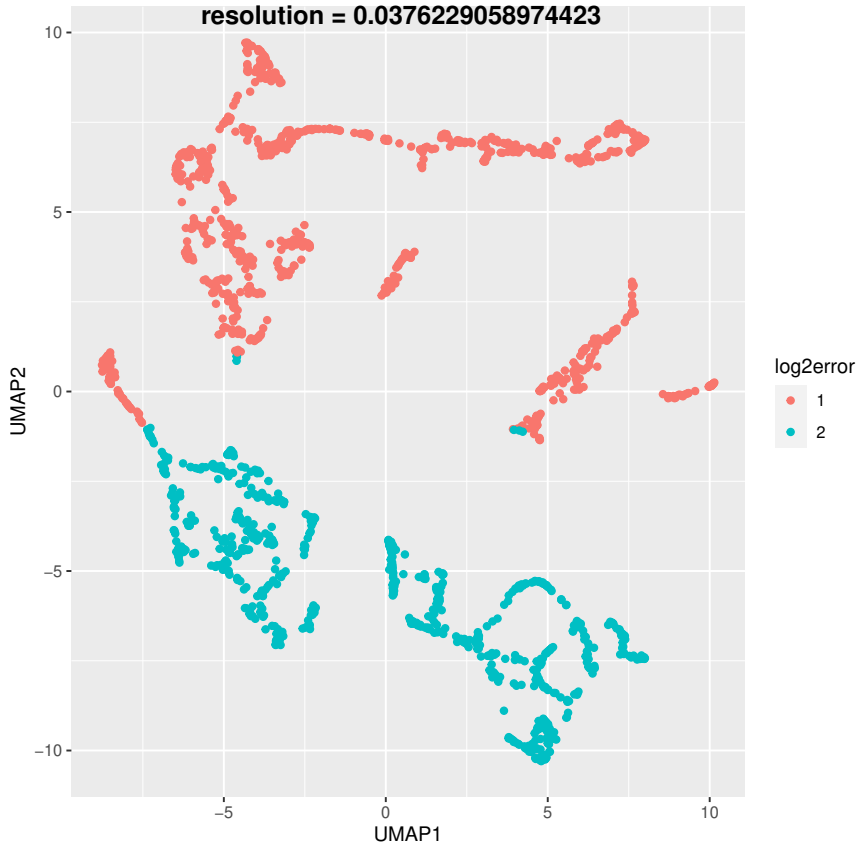

resolution = 1.01124762845226
UMAP2
UMAP1
10
5
0
−5
−10
−5
0
5
10
recall
1
10
11
12
13
14
15
16
2
3
4
5
6
7
8
9

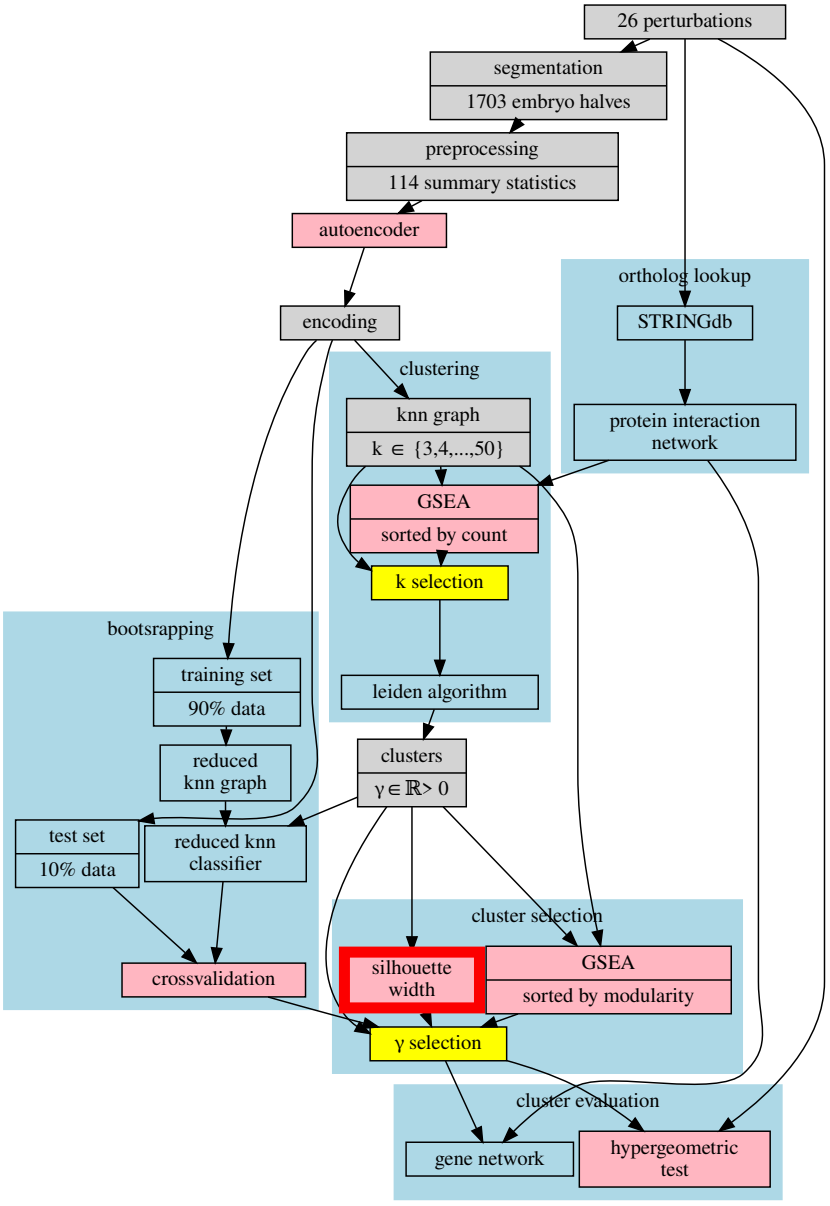


I used this same metric to construct a gene network. If an interaction has positive modularity the genes are considered connected. A recall score can be calculated by comparing this network to the protein interaction network. $-log_2 error$ is calculated by building a kNN classifier for the clusters from a subset of the data and repeating for 1000 permutations.

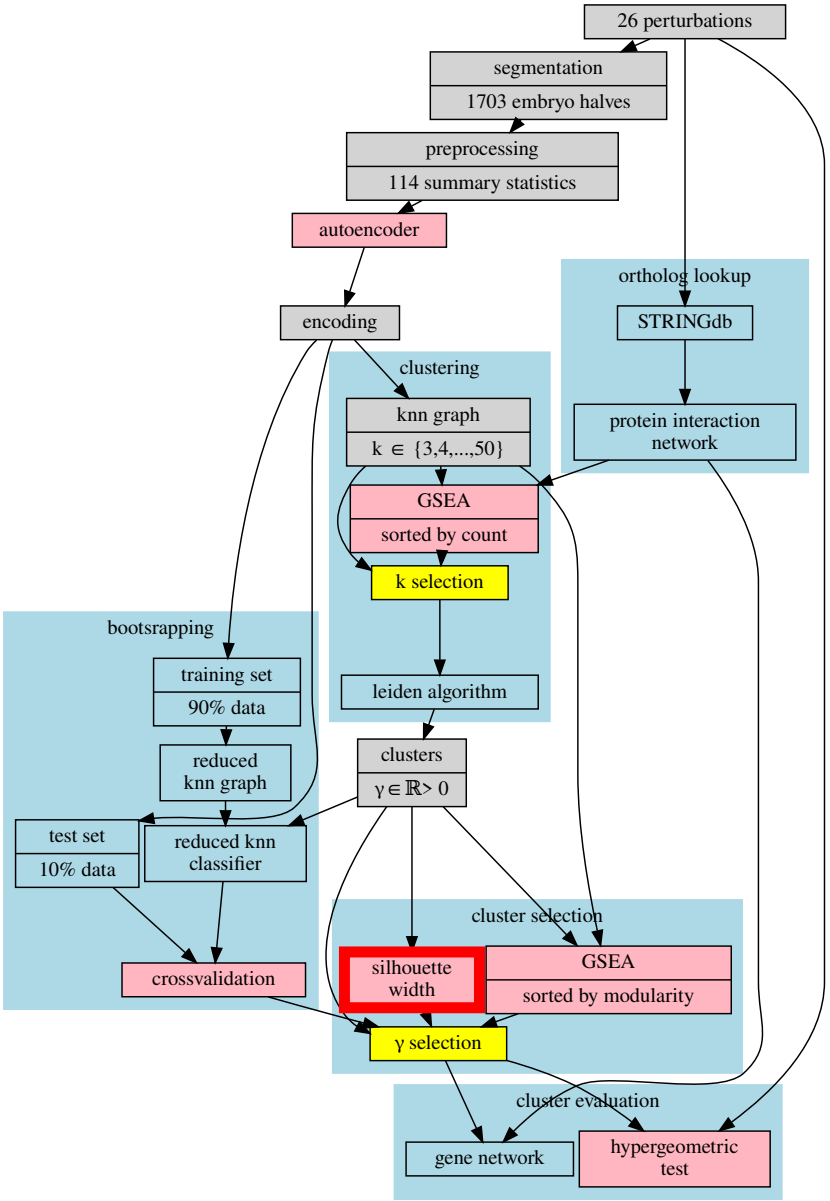

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

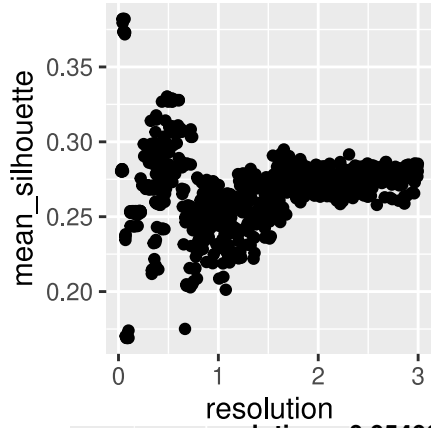

log2error
resolution
10.5
10.0
9.5
9.0
8.5
0
1
2
3

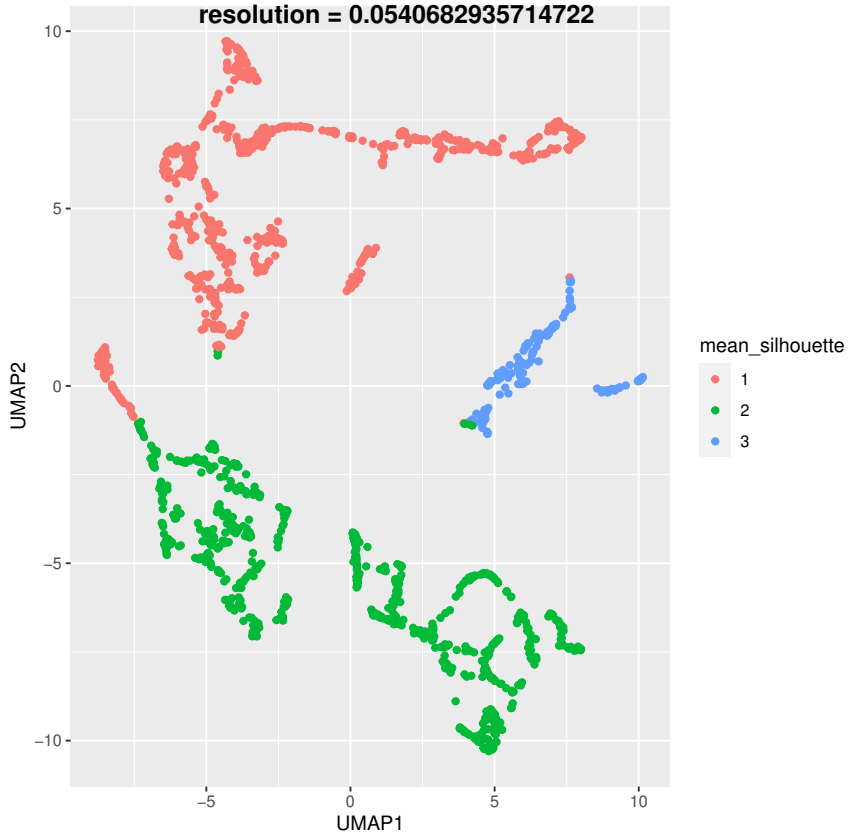

resolution = 0.0376229058974423
UMAP2
UMAP1
log2error
1
2
10
5
0
−5
−10
−5
0
5
10

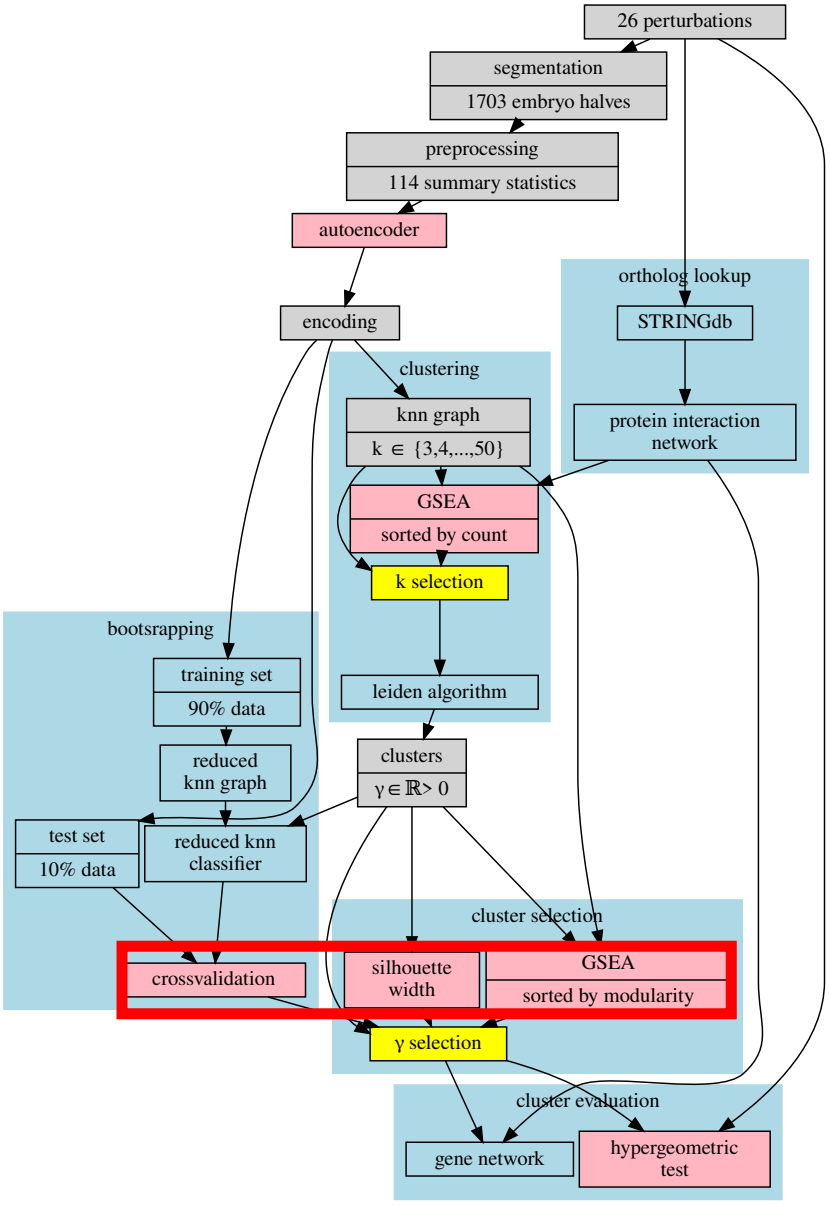


Silhouette score gives the relative distance between points within a cluster compared to distance between points in different clusters.

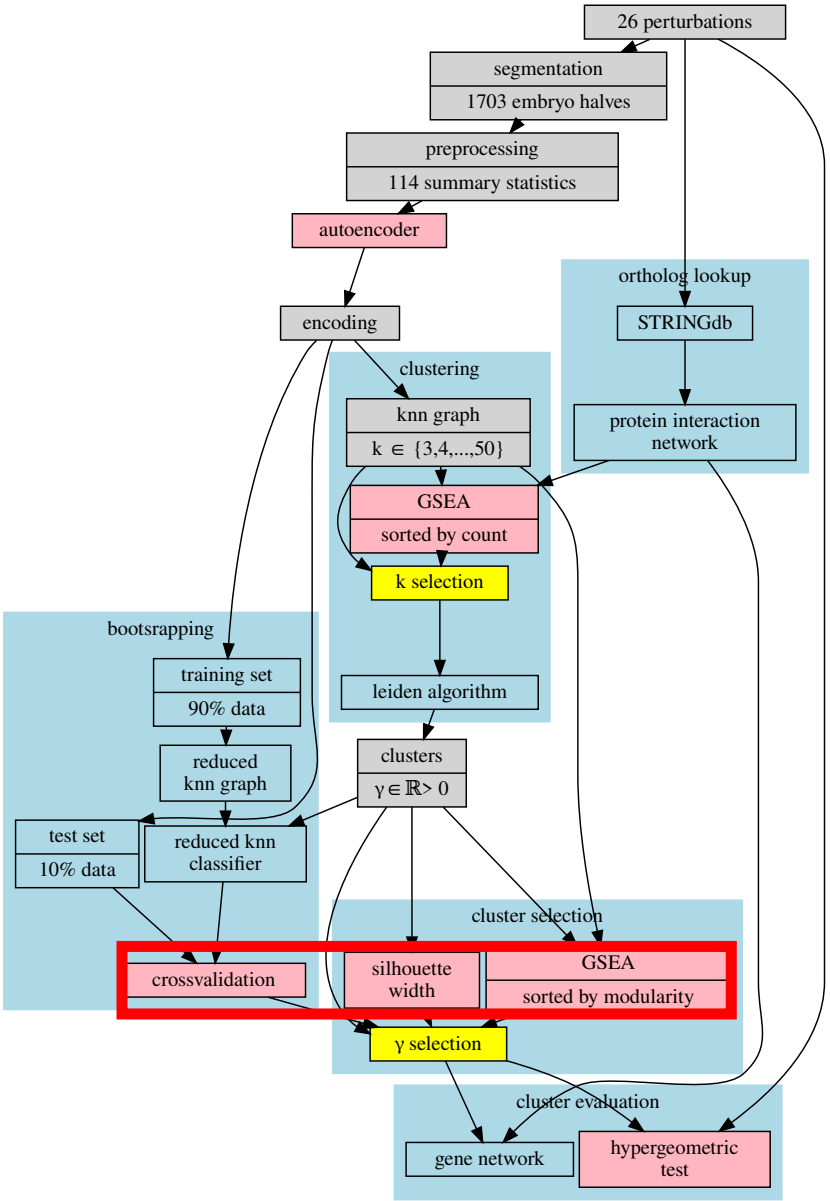

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

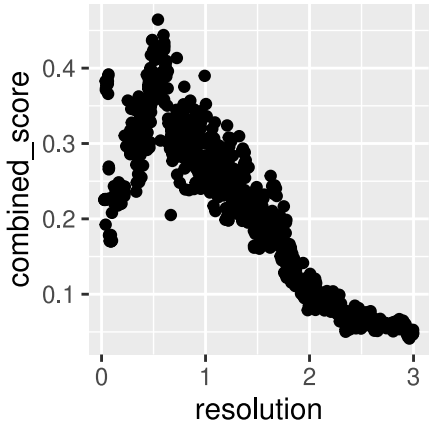

mean_silhouette
0.35
0.30
0.25
0.20
0
1
2
3
resolution

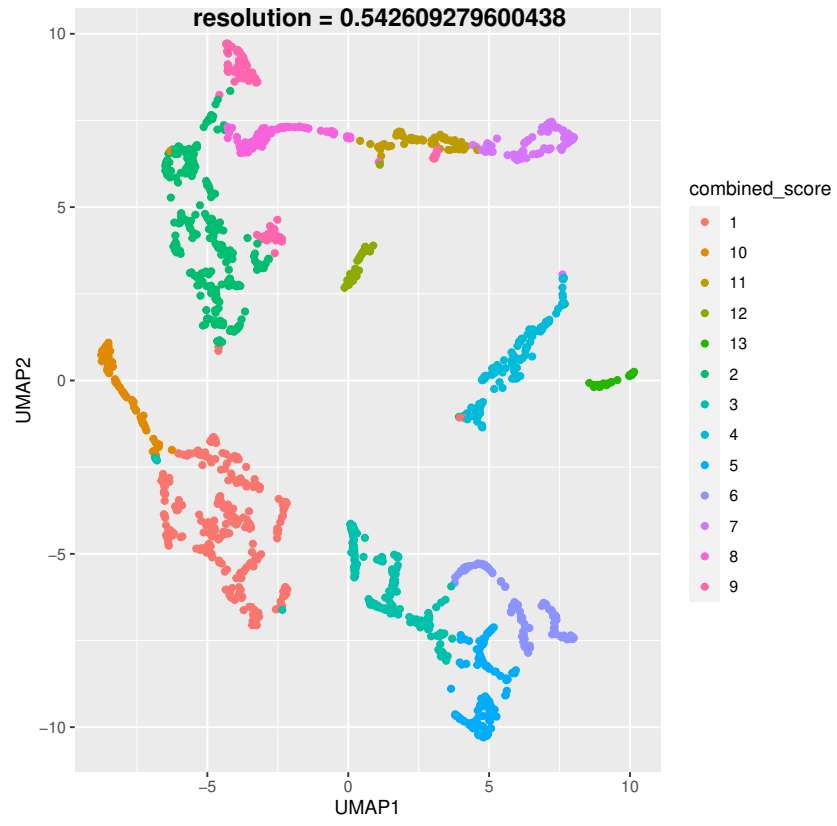

resolution = 0.0540682935714722
UMAP2
10
5
0
−5
−10
−5
0
5
10
UMAP1
mean_silhouette
1
2
3

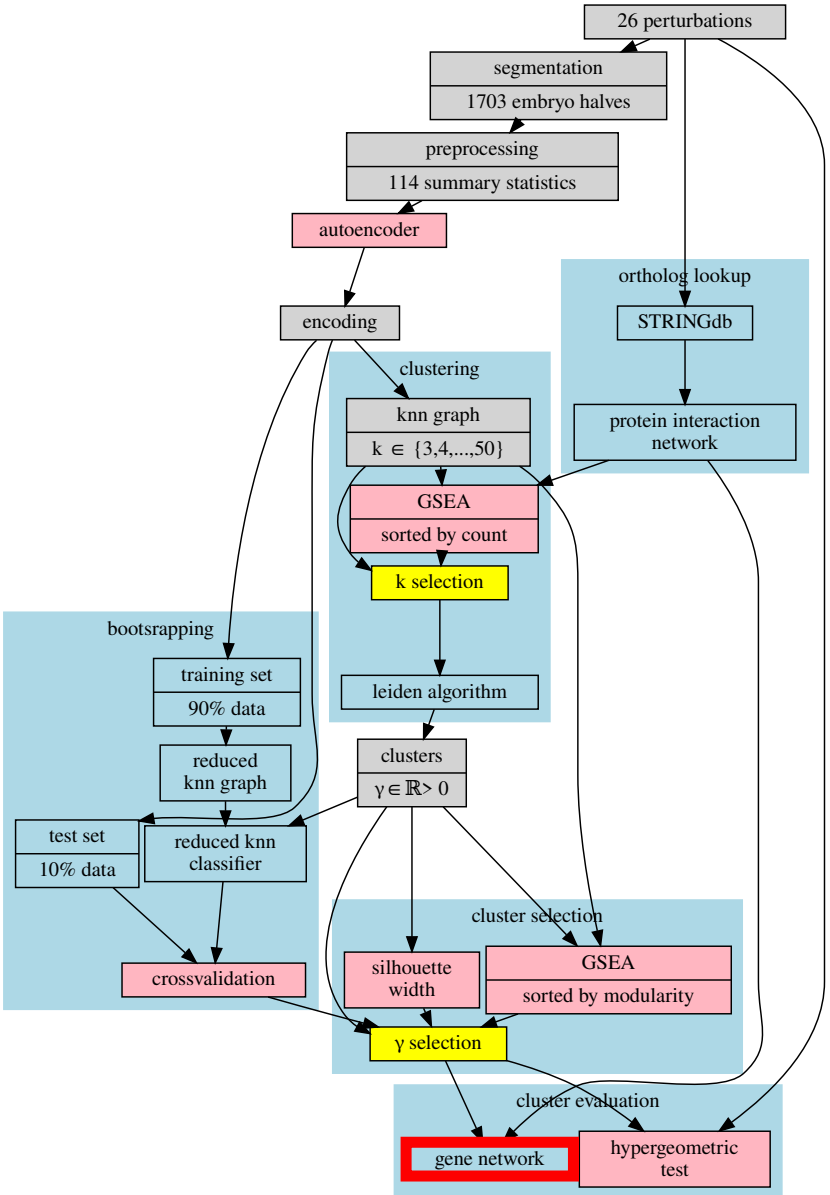


Because there was no clear best metric, I calculated a combined score from $ES \times recall \times silhouettewidth \times -log_2 error$.

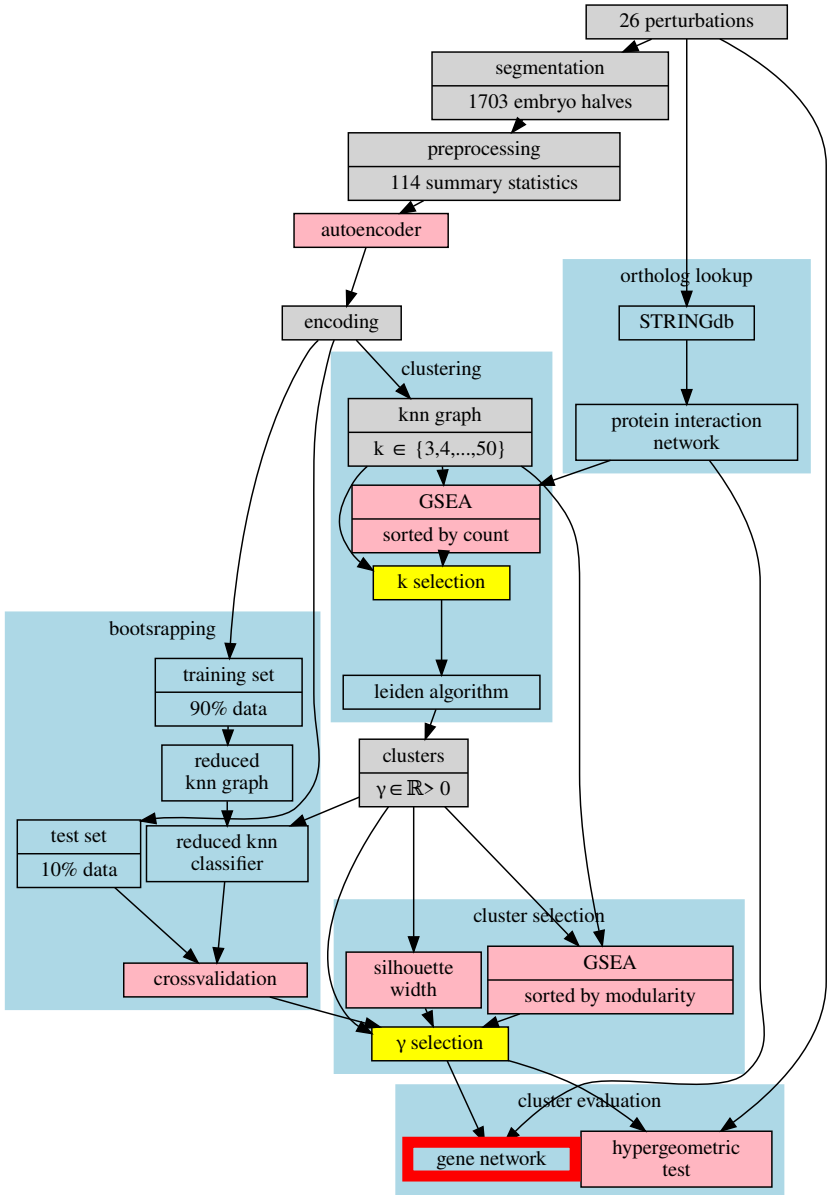

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test

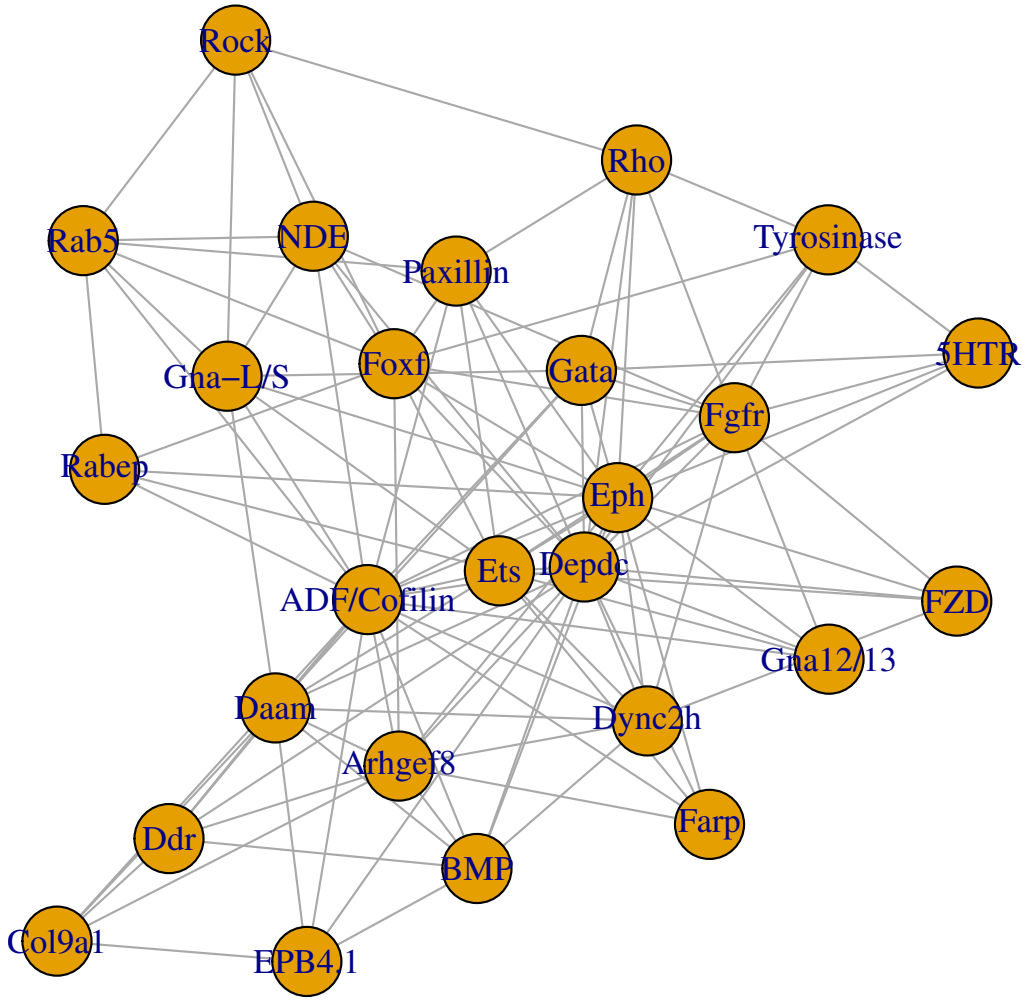

combined_score
0.4
0.3
0.2
0.1
0
1
2
3
resolution

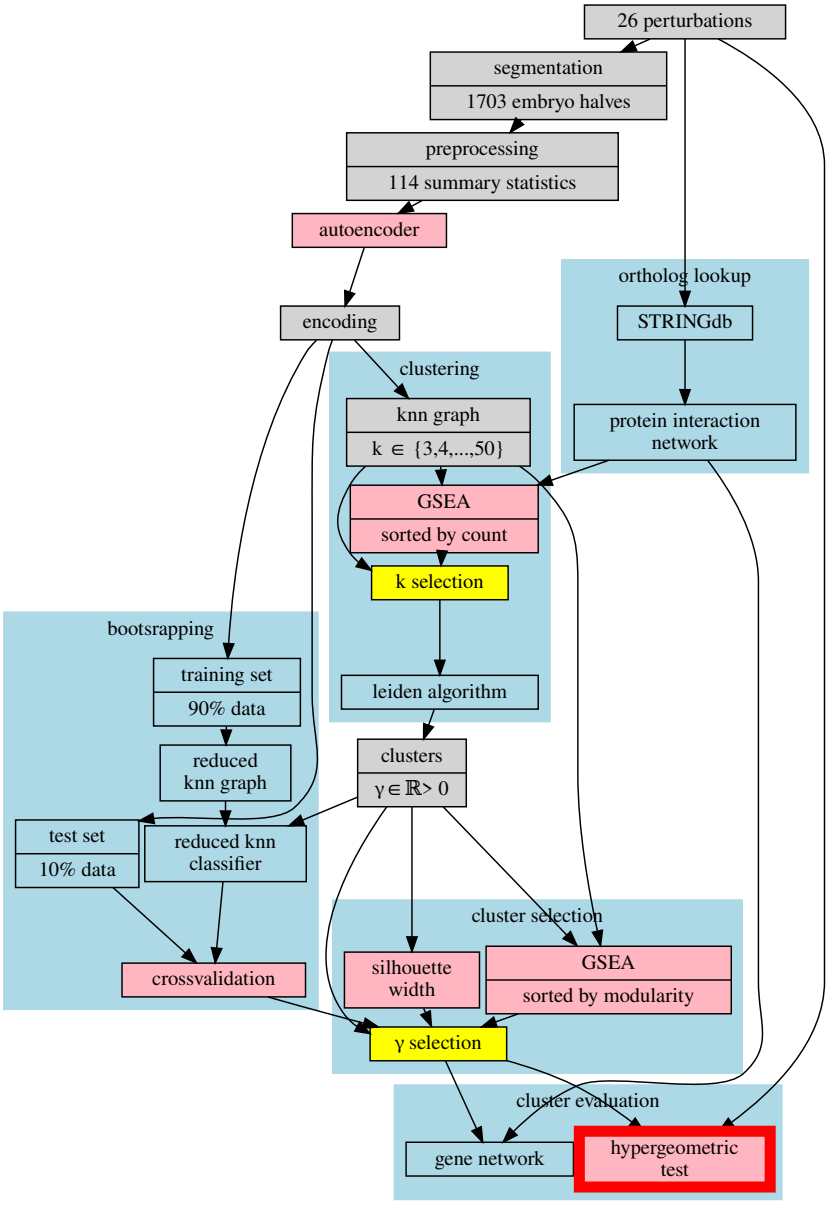

resolution = 0.542609279600438
UMAP2
10
5
0
−5
−10
−5
0
5
10
UMAP1
combined_score
1
10
11
12
13
2
3
4
5
6
7
8
9



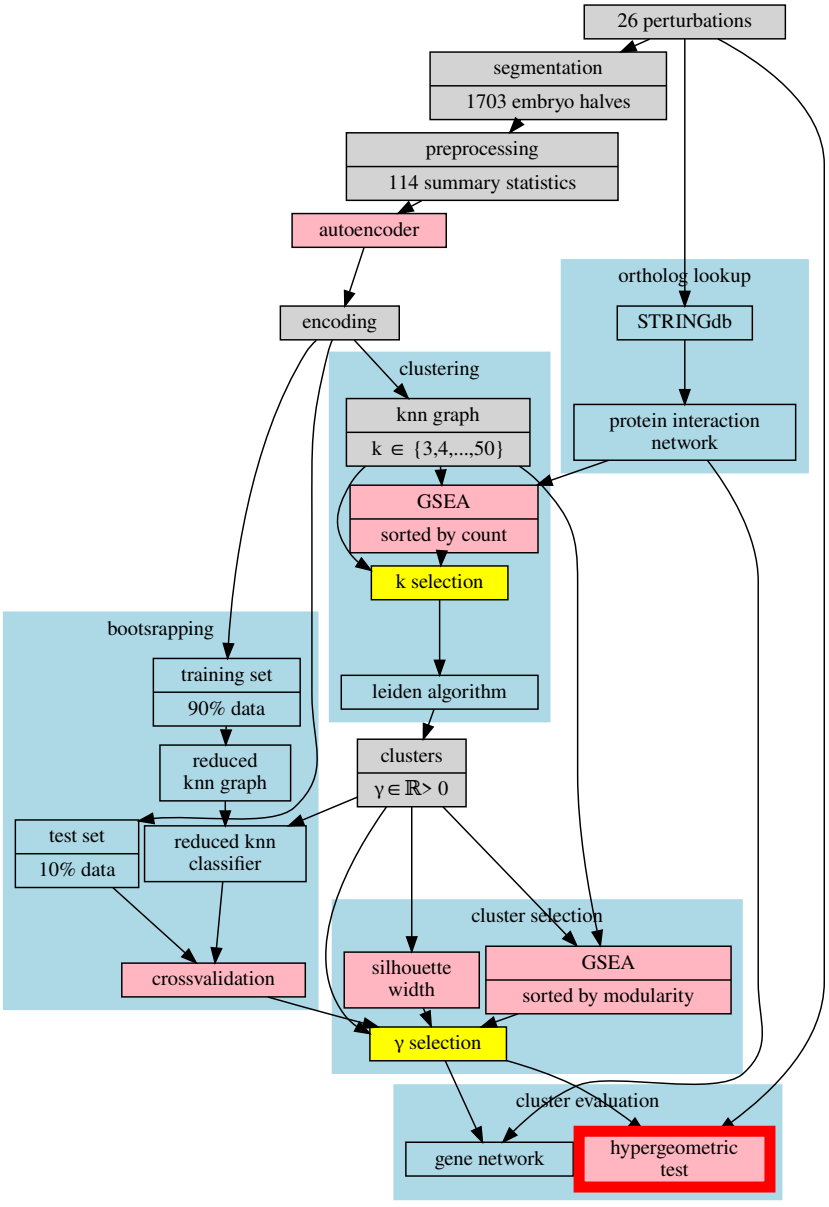
I constructed a network using the top 5 modularity values for each gene.

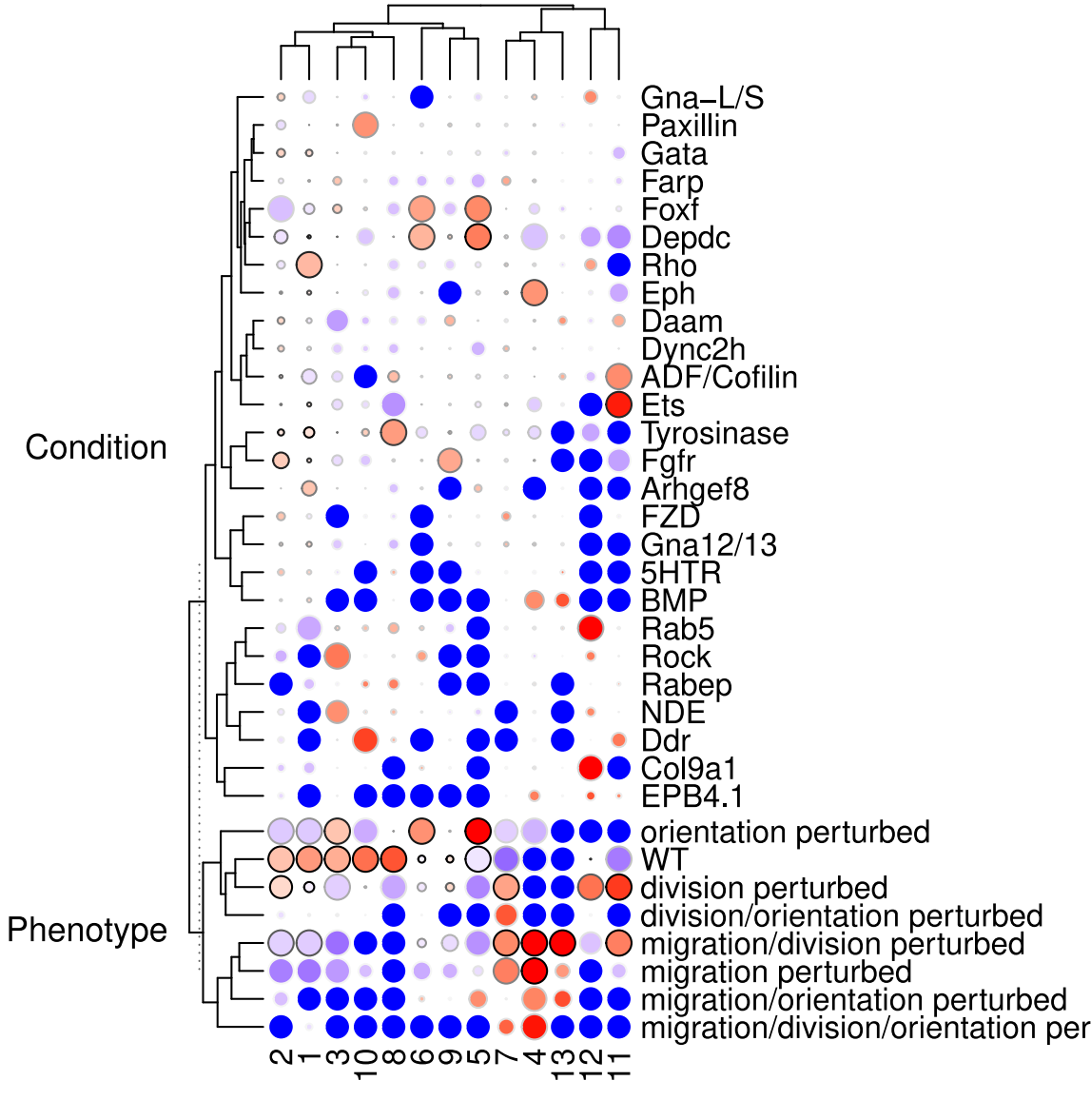

26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
clusters
γ ∈ ℝ> 0
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test


Rock
Rho
Rab5
NDE
Paxillin
Tyrosinase
Gna−L/S
Foxf
Gata
5HTR
Fgfr
Rabep
Eph
Ets
Depdc
ADF/Cofilin
FZD
Gna12/13
Daam
Dync2h
Arhgef8
Farp
Ddr
BMP
Col9a1
EPB4.1



For each cluster I performed hypergeometric tests for enrichment of embryos in each treatment and enrichment of experimenter-identified phenotypes.


26 perturbations
segmentation
1703 embryo halves
preprocessing
114 summary statistics
autoencoder
encoding
ortholog lookup
STRINGdb
protein interaction network
clustering
knn graph
k ∈ {3,4,...,50}
GSEA
sorted by count
k selection
leiden algorithm
bootsrapping
training set
90% data
reduced knn graph
test set
10% data
reduced knn classifier
crossvalidation
clusters
γ ∈ ℝ> 0
cluster selection
silhouette width
GSEA
sorted by modularity
γ selection
cluster evaluation
gene network
hypergeometric test


Condition
Phenotype
Gna−L/S
Paxillin
Gata
Farp
Foxf
Depdc
Rho
Eph
Daam
Dync2h
ADF/Cofilin
Ets
Tyrosinase
Fgfr
Arhgef8
FZD
Gna12/13
5HTR
BMP
Rab5
Rock
Rabep
NDE
Ddr
Col9a1
EPB4.1
orientation perturbed
WT
division perturbed
division/orientation perturbed
migration/division perturbed
migration perturbed
migration/orientation perturbed
migration/division/orientation per
2 1 3 10 8 6 9 5 7 4 13 12 11