Rice gene functional data

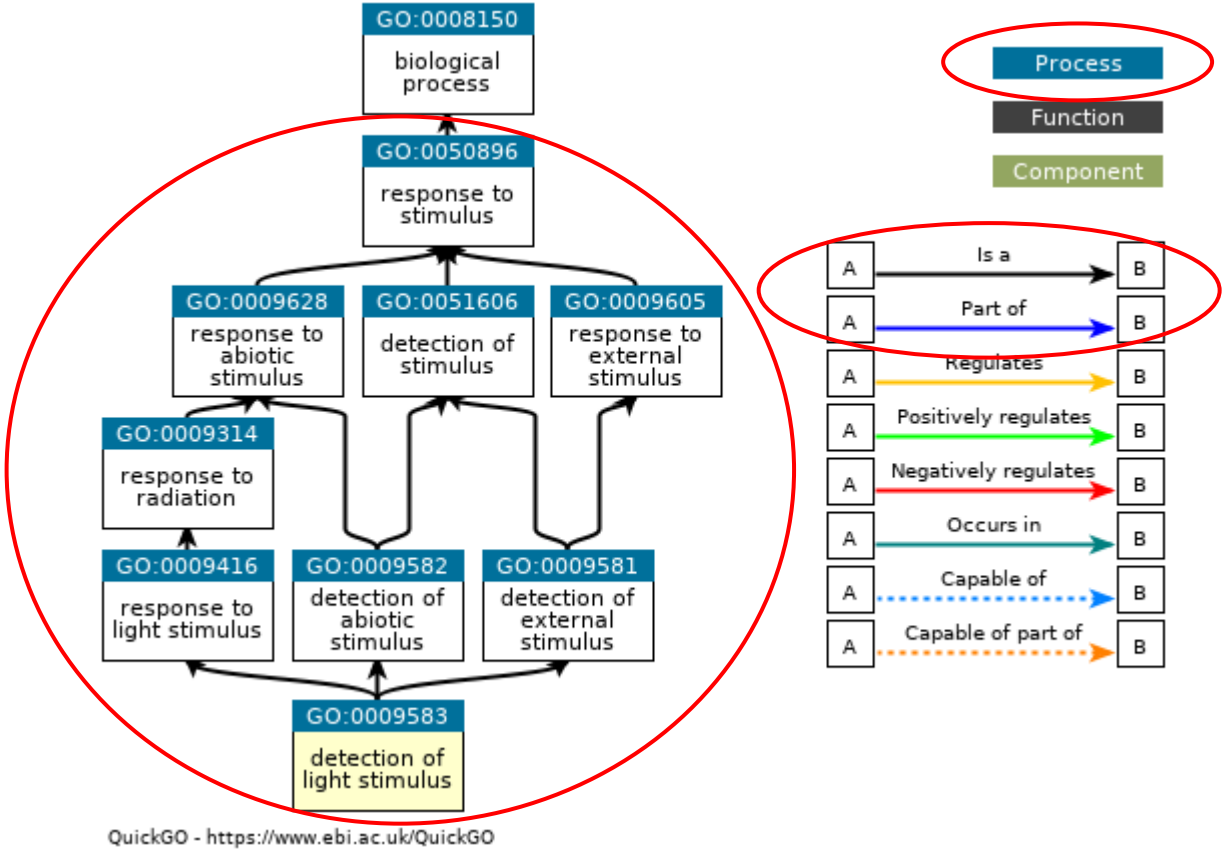
Miguel Romero
Ómicas P5
Sep 18th, 2020

Gene functions

GO:0009583 detection of light stimulus

Ancestor Chart

Function and all its ancestors



Biological processes

Gene functions

Hypothesis

If a gene is associated to a function (biological process) then it is associated to all ancestors of the function.

Total data

Genes	19.663
Co-expression interactions	550.813
Gene-function associations	220.598
Genes with function	9.294
Functions	3.743
Ancestral relations	7.186

Taken from: RAPDB, Oryzabase, Gene2GO, QuickGO, and Gene Ontology.

Term hierarchies

Filtered terms (5<x<=300)	1.797
Terms to predict (including parents)	1.938
Components of DAG	27
Subhierarchies	20
Average number of terms	340
Min-Max number of terms	2-1.996

Term hierarchies

Root	Terms	Pred	Genes	Desc
GO:0044085	121	50	377	cellular component biogenesis
GO:0000003	146	72	648	reproduction
GO:0006796	209	118	1270	phosphate-containing compound metabolic process
GO:0032501	250	120	1043	multicellular organismal process
GO:0032502	290	149	1063	developmental process
GO:0016043	298	140	661	cellular component organization
GO:0051179	325	164	1350	localization
GO:0050896	470	261	3319	response to stimulus
GO:0065007	1027	485	2224	biological regulation
GO:0008152	1463	779	5862	metabolic process
GO:0009987	1996	1025	5900	cellular process

ML approach

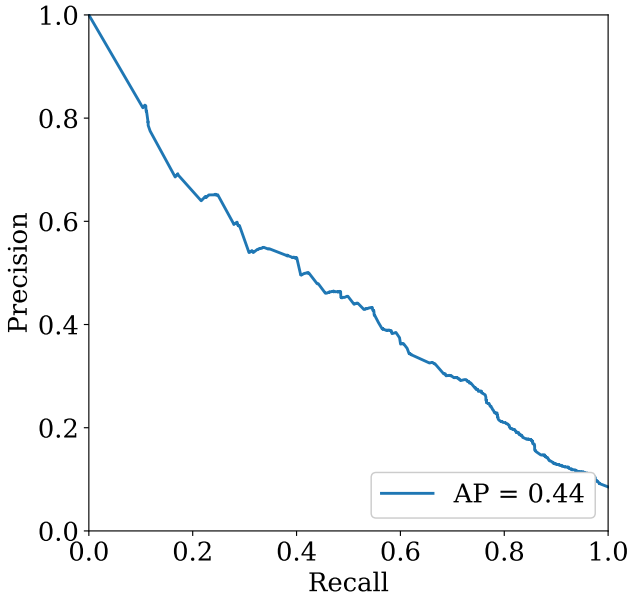
Features

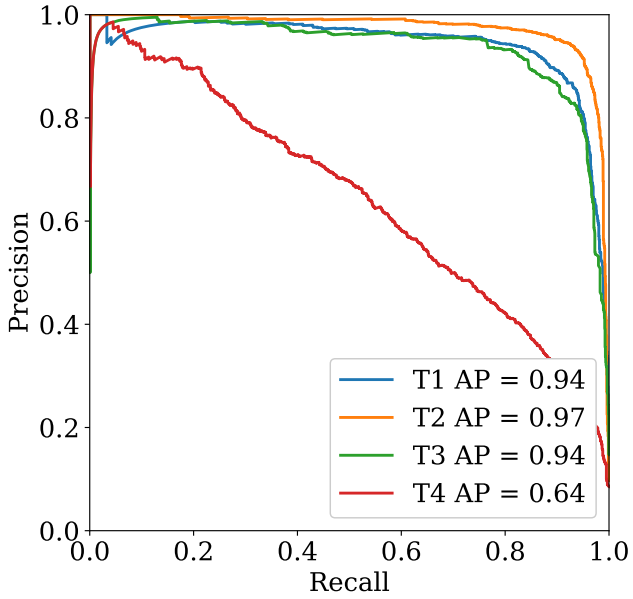
- **Terms:** Gene-function association.
- **TProp:** Topological properties of co-expression network.
- **Emb:** Node2vec embeddings.
- **Emb1D:** Dimension reduction to 1D of Emb.
- **EmbCl:** Clustering from dimension reduction to 2D of Emb.

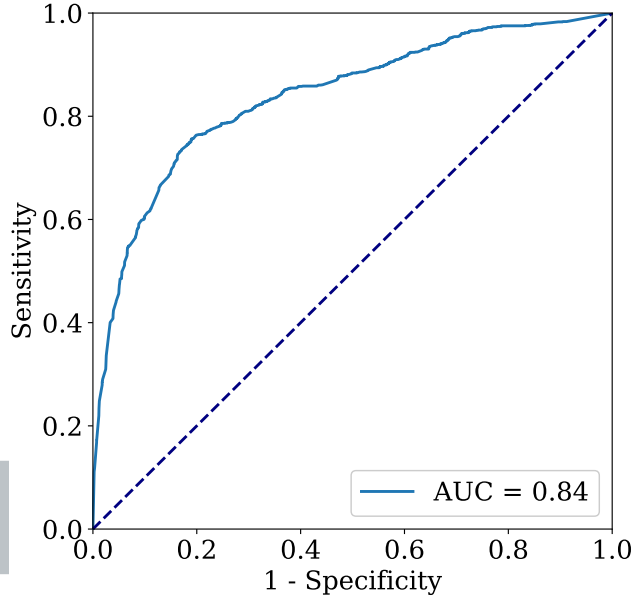
Datasets

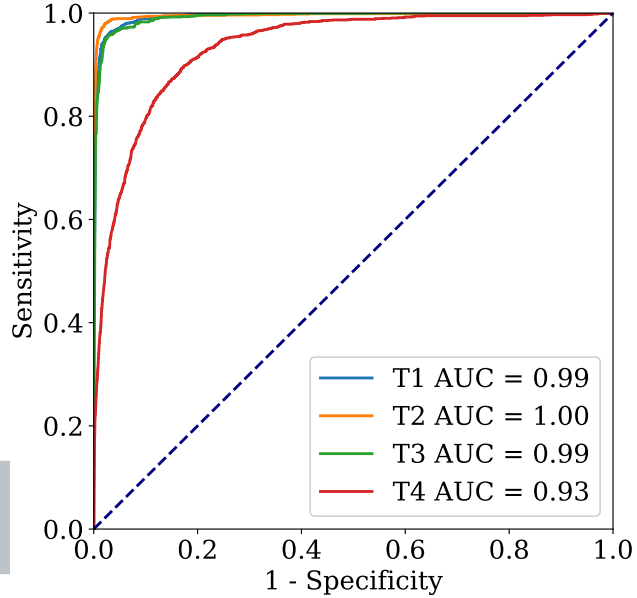
- **T1:** Terms, TProp, Emb1D
- **T2:** Terms, TProp, Emb, Emb1D, EmbCl
- **T3:** Terms, TProp
- **T4:** TProp, Emb1D, EmbCl

HBN vs. ML: GO:0044085

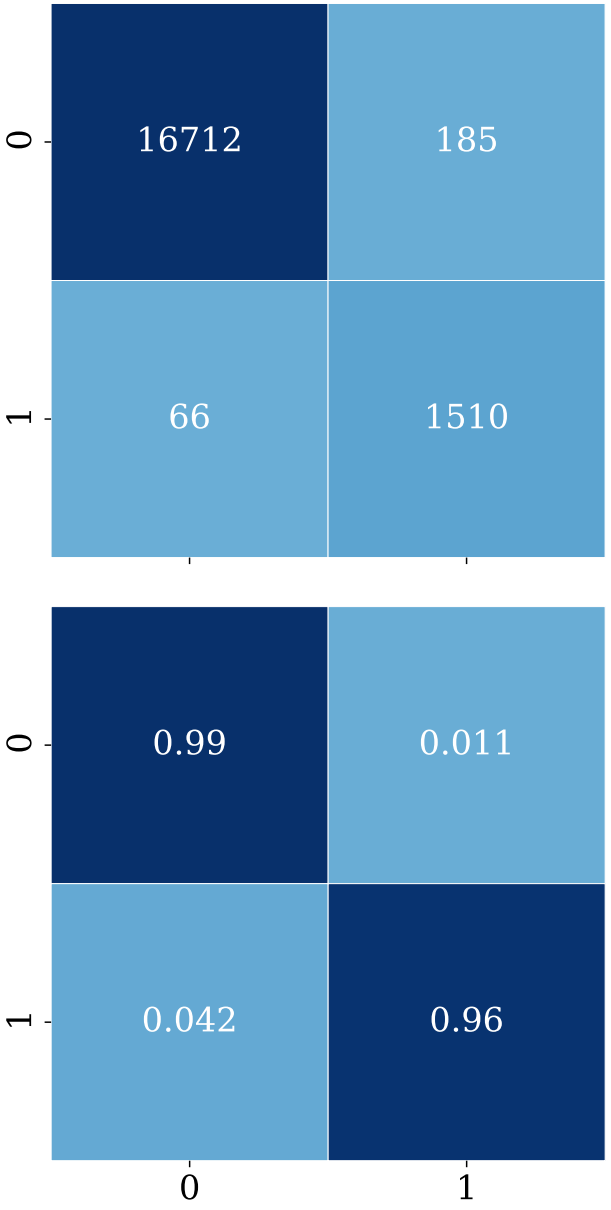


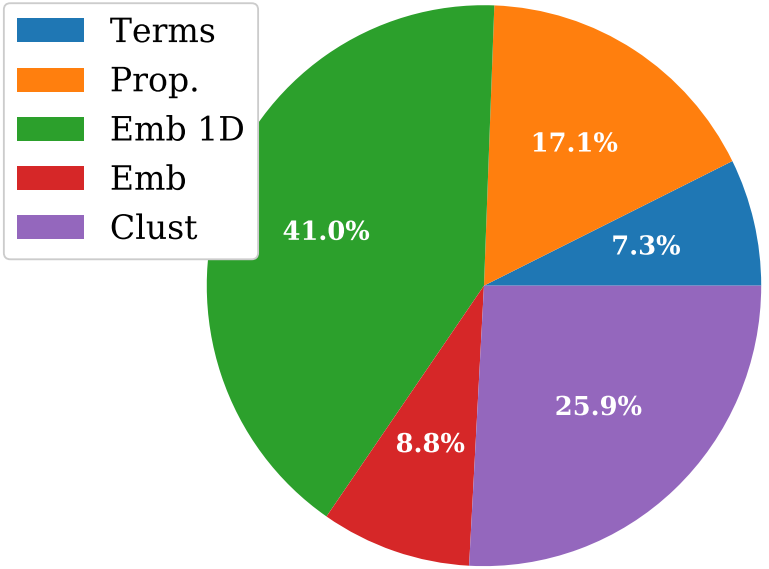


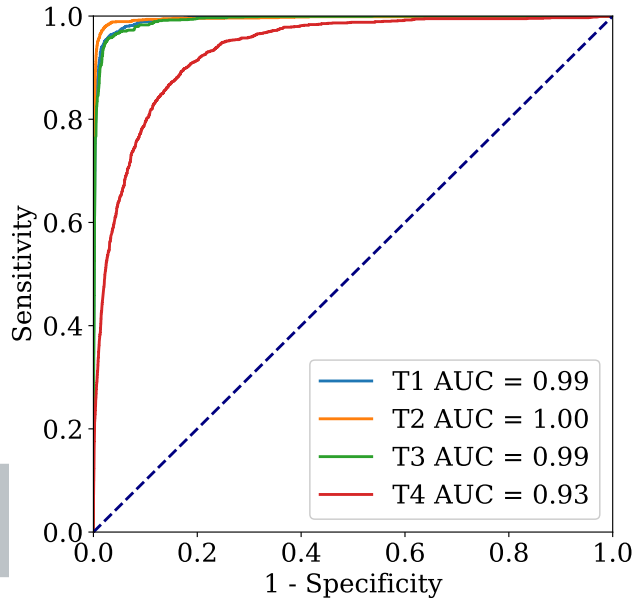




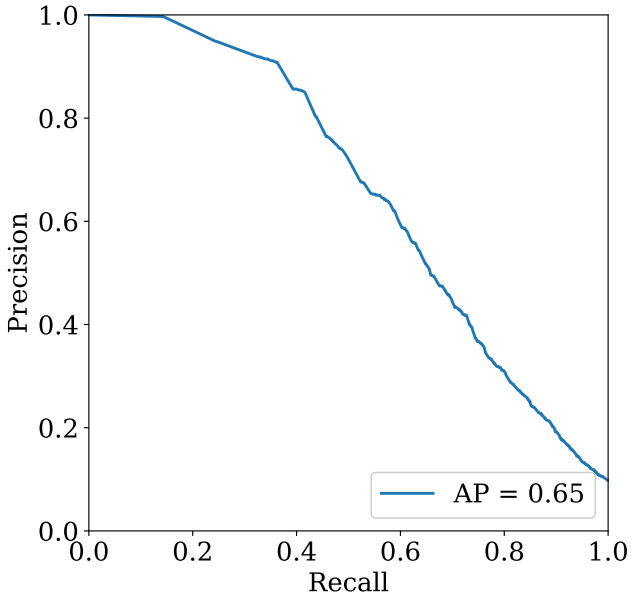
HBN vs. ML: GO:0044085

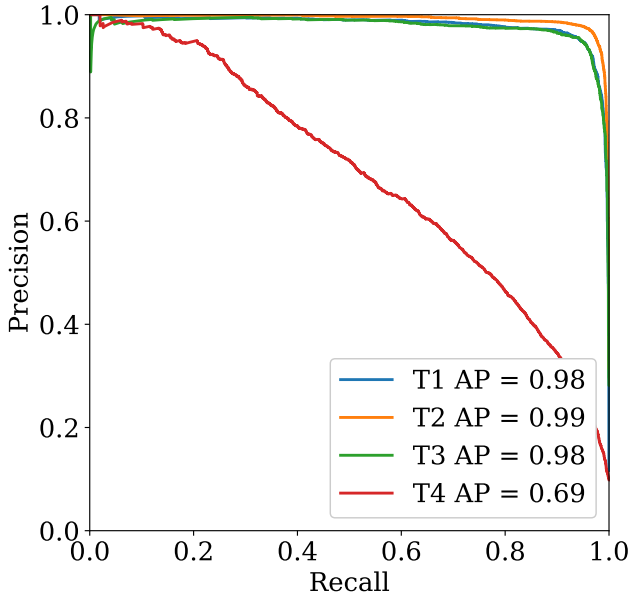


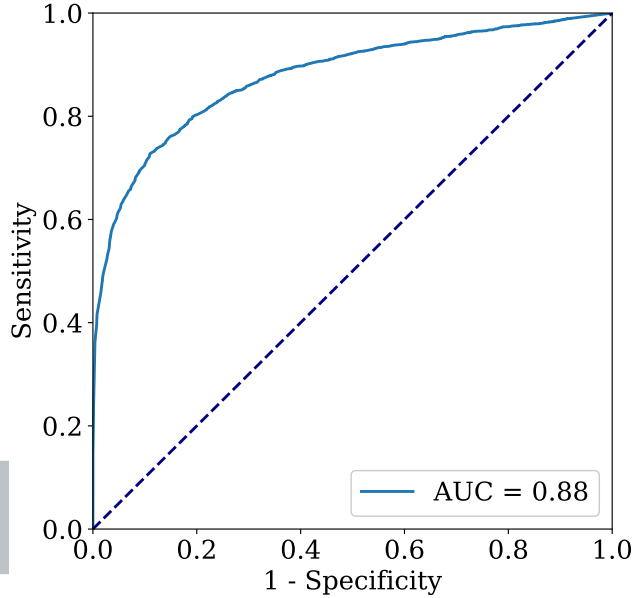


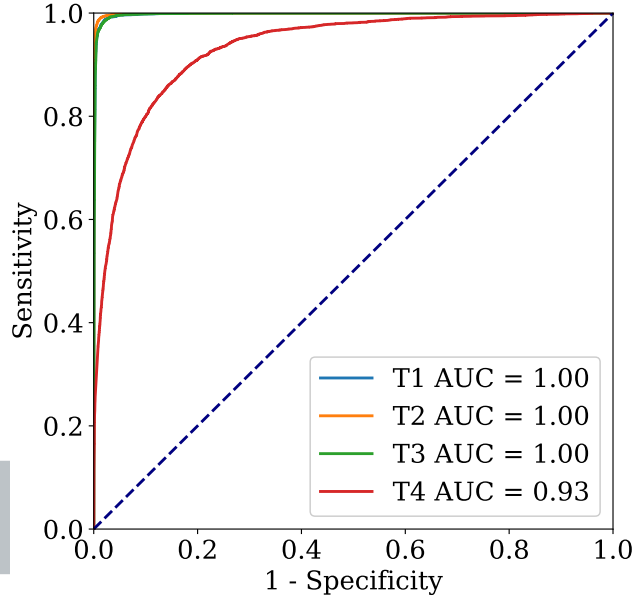


HBN vs. ML: GO:0000003

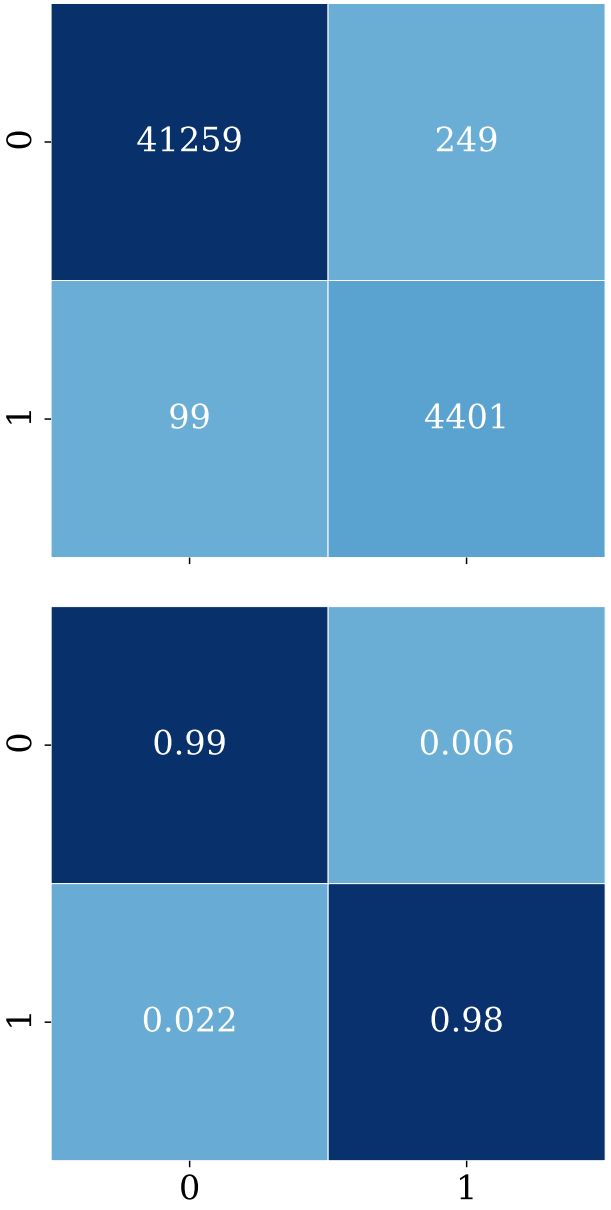


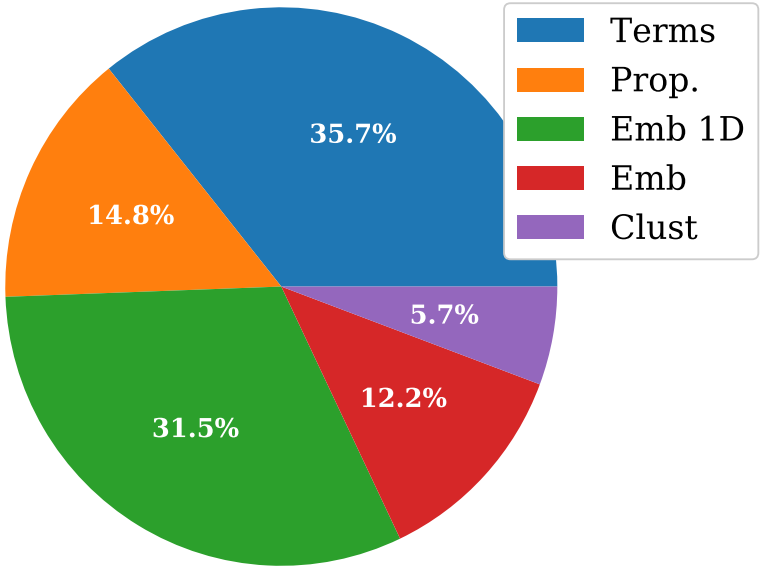


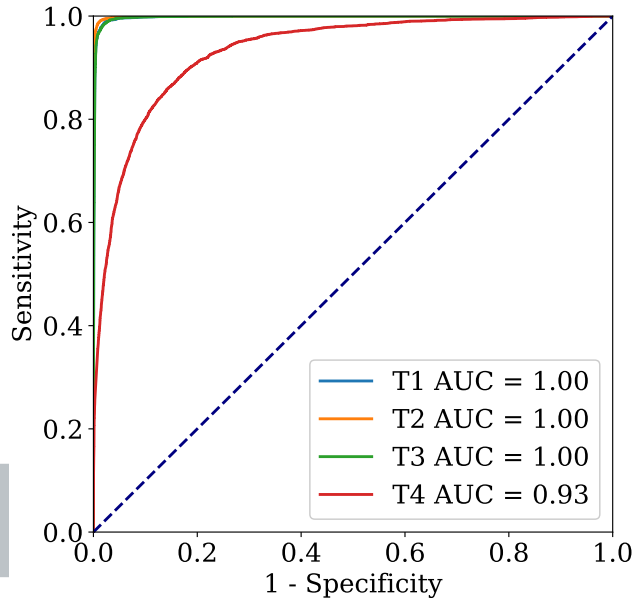




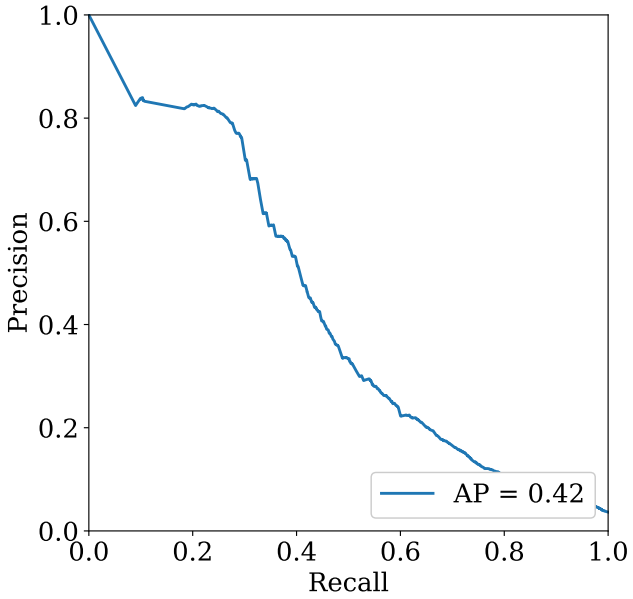
HBN vs. ML: GO:0000003

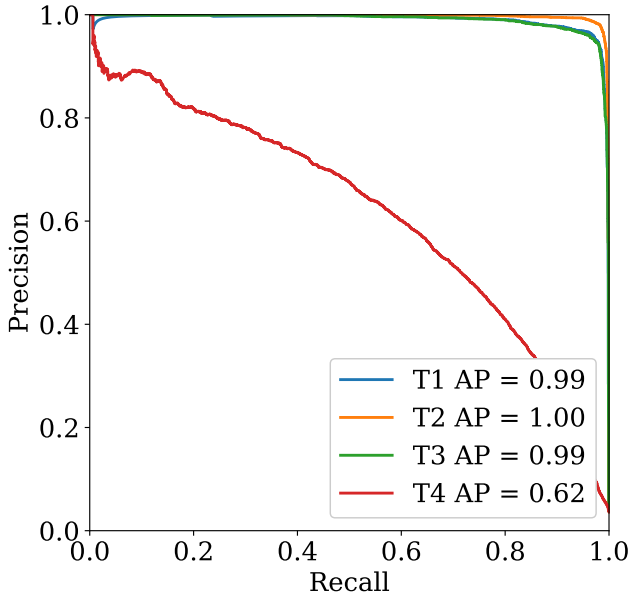


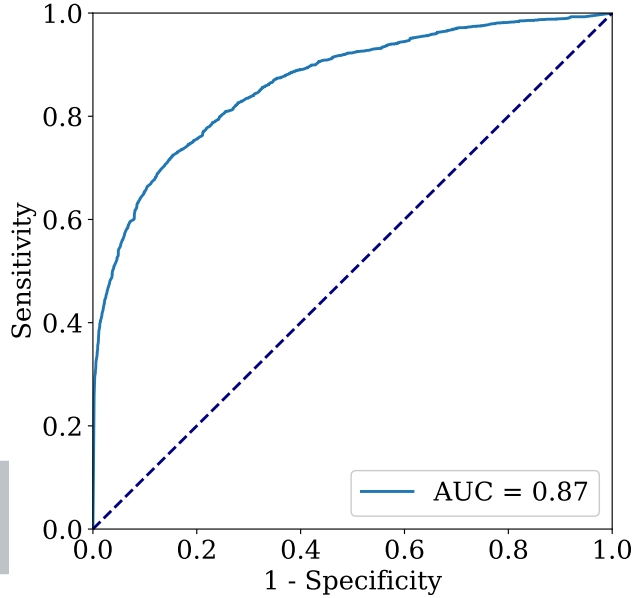


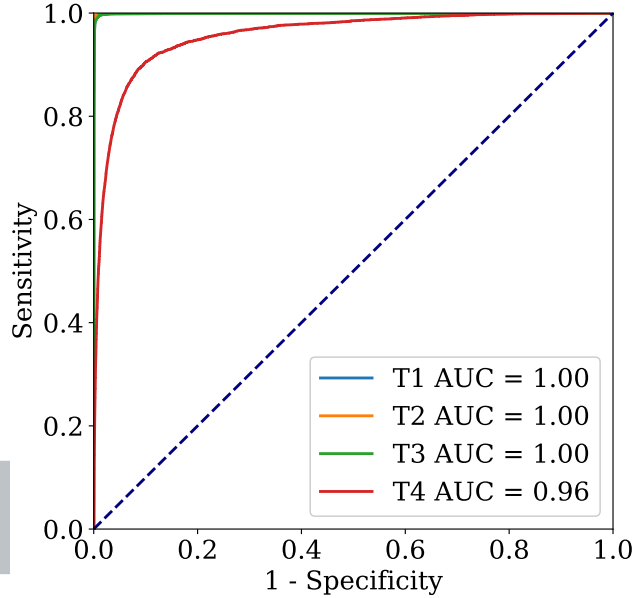


HBN vs. ML: GO:0006796

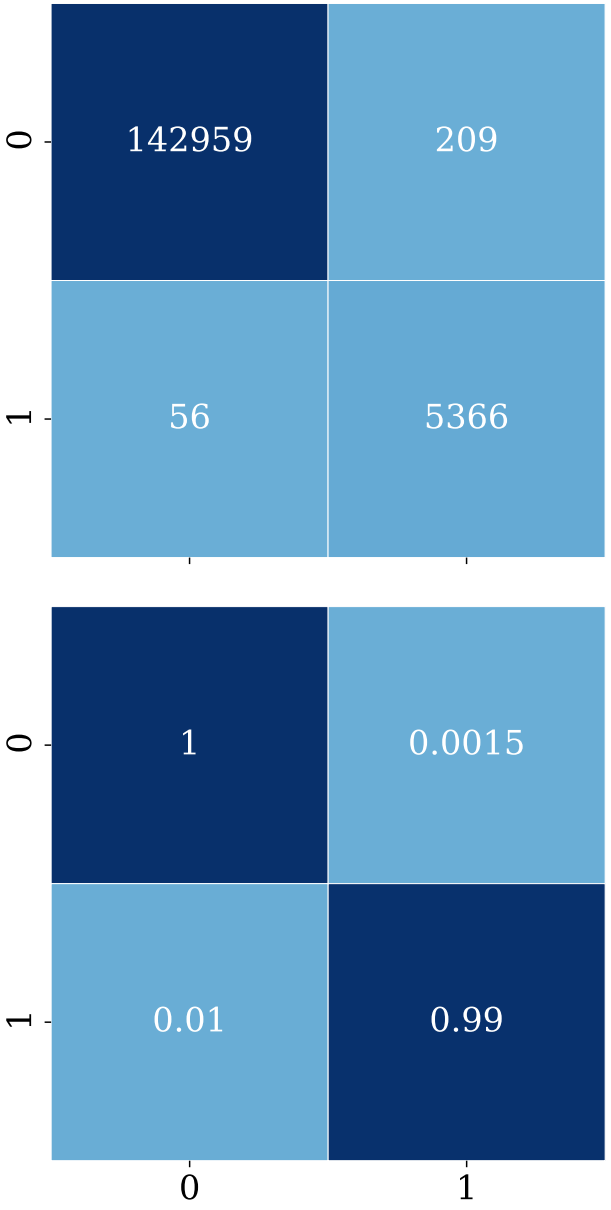


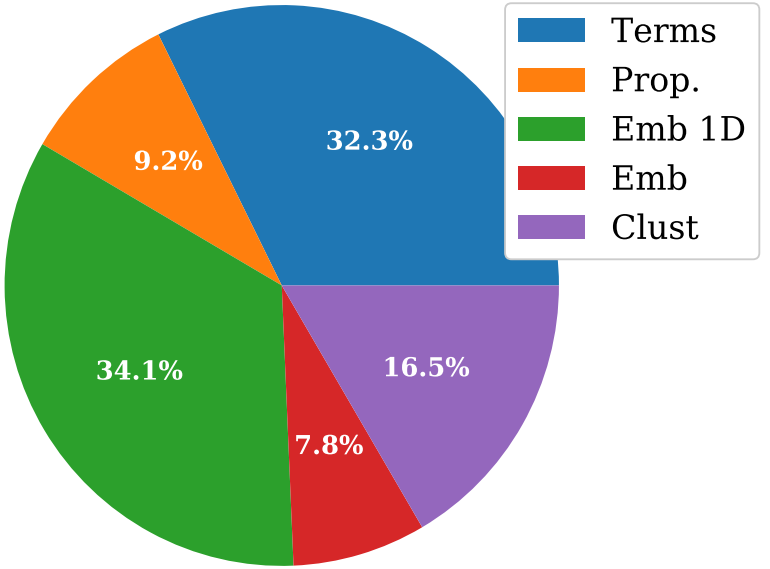


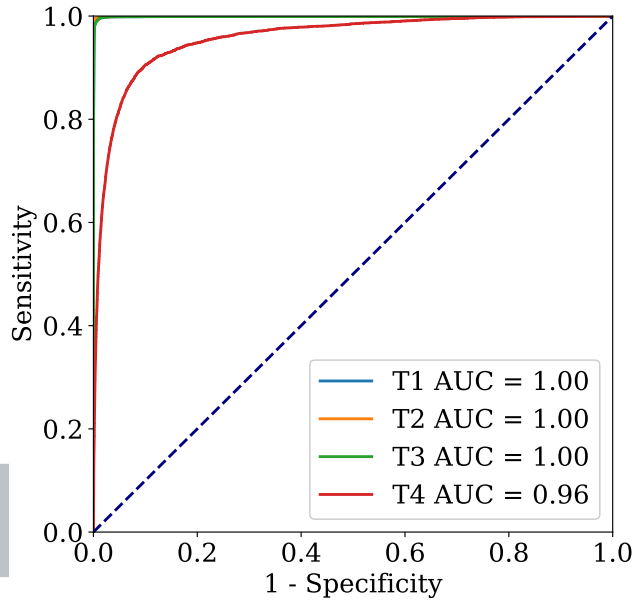




HBN vs. ML: GO:0006796







Improve HBN algorithm

Before

- Prediction is made per gene and term.
- Prediction of each gene and hierarchy is independent.
- Only ROC AUC and average precision metrics are computed.

After

- Run independent predictions in parallel, this will drastically reduce execution time.
- Include confusion matrix and log loss metrics.

Change ML prediction strategy

Before

- Whole hierarchy is considered in prediction
- Target term is removed form dataset
- True labels of genes are considered in the dataset

After

- Consider hierarchy from top to bottom in a BFS manner
- No need to remove term
- Consider predicted labels of genes in the dataset

Change ML prediction strategy

Before

- The size of the dataset is the same for each term in the hierarchy
- Inconsistencies are fixed at the end

After

- Improve efficiency of the algorithm, since dataset size is reduced
- Fix inconsistencies along the way

Probles with experiments

- Hydra used to be busy.
- Some nodes are down very often.
- When there are more processes in the same node the performance reduce drastically. Execution time change from 2s to more than 60s per iteration.
- Processes are killed unexpectedly.

To do

- Experiments with HBN model on Hydra.
  - Run in parallel and add more metric, i,e, log loss and confusion matrix (next).

- Experiments with ML model on Hydra (ongoing).
  - Use graph embeddings with *node2vec* (ongoing).
  - Change training and prediction strategy (next).
  - Analyze feature importance (next).
  - Use graph embeddings with *GCN* (next).

- Compare performance of the models.