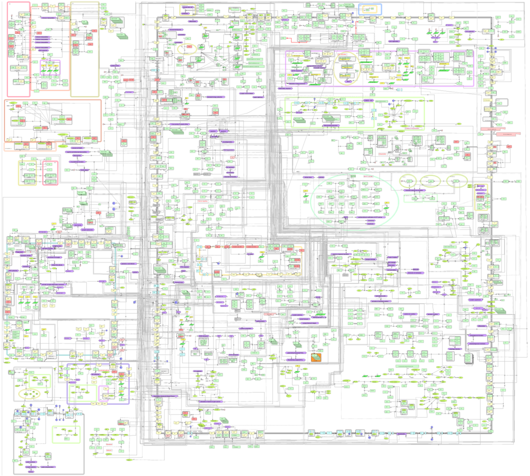
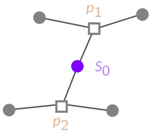
Node Duplication in Disease Maps using Graph Neural Networks

Colloquium



- Can always make layout task easier by duplicating nodes with degree ≥ 2
- But which nodes can be duplicated s.t. network information remains faithful?

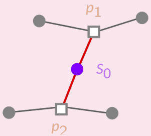
TODO



Single species alias may be connecting multiple processes

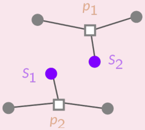
Path (p_1, S_0, p_2) is semantically meaningful (**true connectivity**)

$\rightsquigarrow$ S_0 must not be duplicated



Path (p_1, S_0, p_2) is not meaningful (implies **false connectivity**)

There should be no paths implying false connectivity
$\rightsquigarrow$ S_0 should be **duplicated**



e.g. due to unrelated roles of S_0 in p_1, p_2, not stoichiometrically linked, unimportant byproduct

TODO

Objective 1

Assess whether a given species alias implies false connectivity (and should thus be duplicated)

here: depends on context etc.?

Objective 2

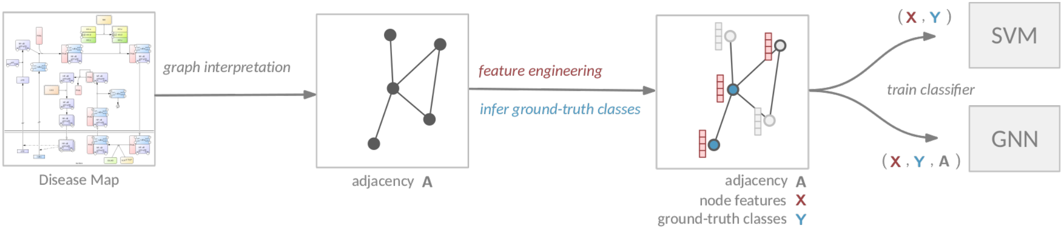
Determine number of duplicates and attachment of edges

Some previous approaches would rely on **node centrality scores**
high centrality ⇝ heterogeneous neighbourhood ⇝ false connectivity

- node degree [**?**, **?**]
- eigenvector centrality [**?**]
- communities (modularity)
  - contribution to modularity if node removed [**?**]
  - based on intra- & inter-community degrees [**?**]
- communities (semantic)
  - cellular compartment [**?**]
  - pathway annotation [**?**, **?**, **?**]

Objective

Given expert decisions, train an ML model to predict node duplication.



Disease Map

graph interpretation

adjacency **A**

feature engineering

infer ground-truth classes

adjacency **A**
node features **X**
ground-truth classes **Y**

(**X** , **Y**)

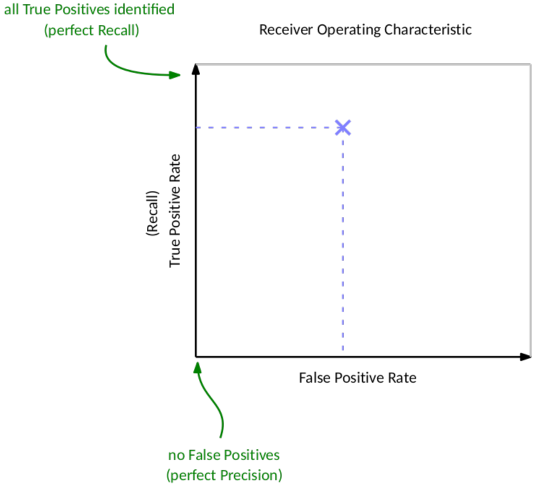
train classifier

(**X** , **Y** , **A**)

SVM

GNN

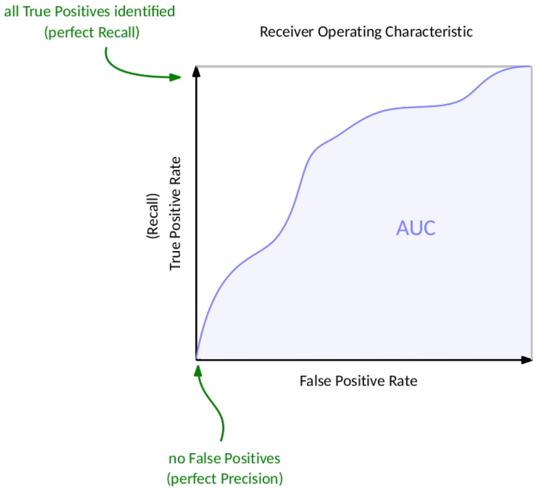
- To compare classifiers, we need an **unbiased performance measure**
- Classifiers yield a **confidence score** in $[0, 1]$ for a given example
- Obtain concrete classification by setting a **decision threshold**
- **True Positive Rate** (TPR): $^{\#\text{ true positives}}/_{\#\text{ actually positive}}$
- **False Positive Rate** (FPR): $^{\#\text{ false positives}}/_{\#\text{ actually negative}}$
- Usually a tradeoff, choice depends on use-case
  - Accept only few high-confidence predictions → low FPR, but also low TPR (Recall)
  - Lower decision threshold → increase TPR at cost of increased FPR



all True Positives identified (perfect Recall)

no False Positives (perfect Precision)

- Concrete choice of threshold yields binary classification and TPR, FPR



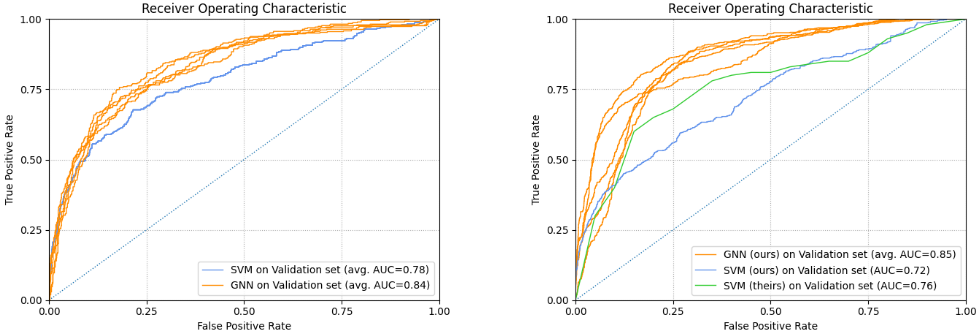
all True Positives identified (perfect Recall)

no False Positives (perfect Precision)

Plot TPR and FPR as function of decision threshold

- Plot TPR, FPR as function of decision threshold $\rightsquigarrow$ **ROC curve**
- Useful properties:
  - Show overall behaviour with respect to variable threshold
  - Insensitive to class distribution
  - Insensitive to error costs

Block Example



(a) (AlzPathwayReorg → PDMap)

(b) (AlzPathwayReorg → ReconMap)

foo

- foo bar baz flubble qox cazinga
- flofola kinorrat ewusa a