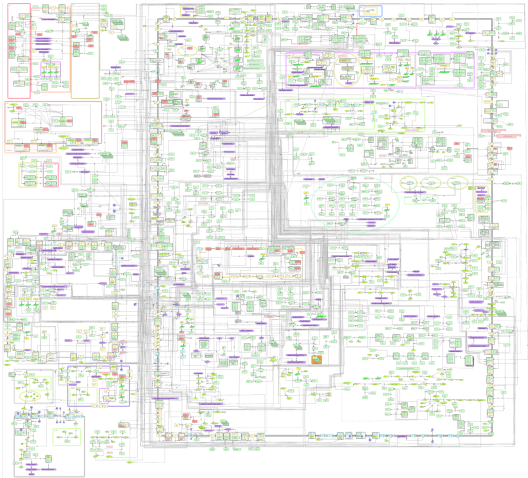
Node Duplication in Disease Maps using Graph Neural Networks

Colloquium

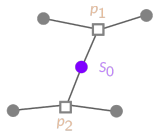
- In preliminary layouts of disease maps, a species alias may be connected to many different processes
- Unclear whether connections are meaningful or merely artifact of creation process
- Faithful network representation should not have such connections $\rightsquigarrow$ duplicate some nodes
- Decision which nodes may be duplicated is not trivial $\rightsquigarrow$ consider ML model trained on expert decisions
- ...

(definition of species, species alias, process, ...) (DMs interpreted as networks) (details on graph interpretation?)



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

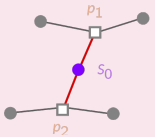
TODO



Single species alias may be connecting multiple processes

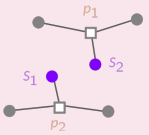
Path (p_1, S_0, p_2) is semantically meaningful (**true connectivity**)

$\rightsquigarrow$ 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

distinction between these two tasks...?

Some previous approaches would rely on **node centrality scores**
high centrality $\rightsquigarrow$ heterogeneous neighbourhood $\rightsquigarrow$ 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 1

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

- No clear requirements or guidelines (yet)
- Previous work relies on heuristic rules
- Decision potentially depends on biological domain knowledge.
- Decision depends on *context* (neighbourhood) of given species alias

$\rightsquigarrow$ Try to learn rules from examples provided by domain expert
$\rightsquigarrow$ Provide information on network structure

Given expert decisions, train a ML model for **supervised node classification** to predict node duplication.

1. "expert decisions"?
2. "supervised node classification"?

"Expert decisions"? (repeat first part of later slide here)

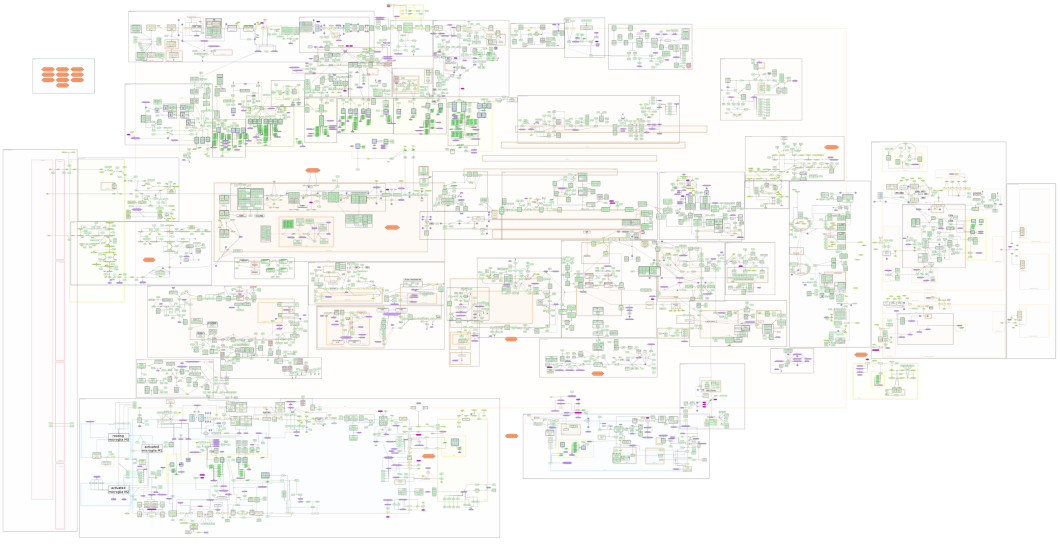
(step through screenshots of snapshots...)

TODO

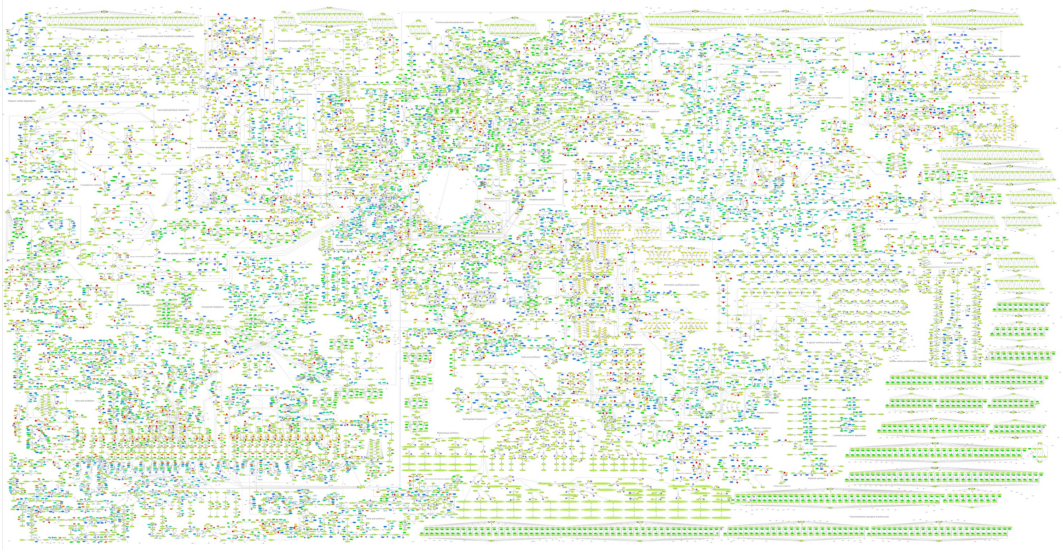
- **AlzPathway** is a disease map that describes signalling pathways related to Alzheimer's Disease
- Recently received additional curation of layout, including duplication of nodes
- Snapshots of intermediate progress were saved (**reorganisation steps**)

- Such reorganisation steps are hard to obtain in practice
- For any disease map, can create a single "step" by comparing it to its **collapsed** version

TODO



Parkinson's Disease Map (PDMap) describes major pathways involved in pathogenesis of Parkinson's Disease

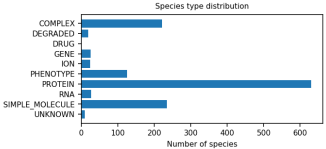


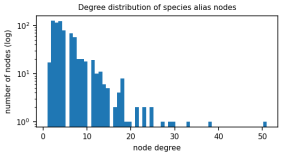
ReconMap [?], a visual representation of the *Recon 2* [?] GSMM

TODO datasets used

AlzPathwayReorgLast (collapsed)

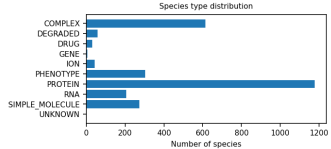
1315 species aliases, 982 reactions
719 (54.68%) for prediction
277 (38.53%) to-be-duplicated

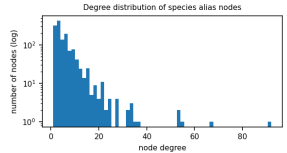




PDMap19 (collapsed)

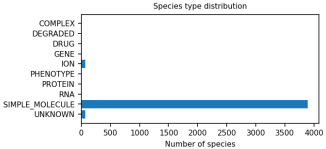
2706 species aliases, 2661 reactions
1389 (51.33%) for prediction
393 (28.29%) to-be-duplicated

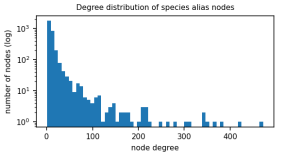




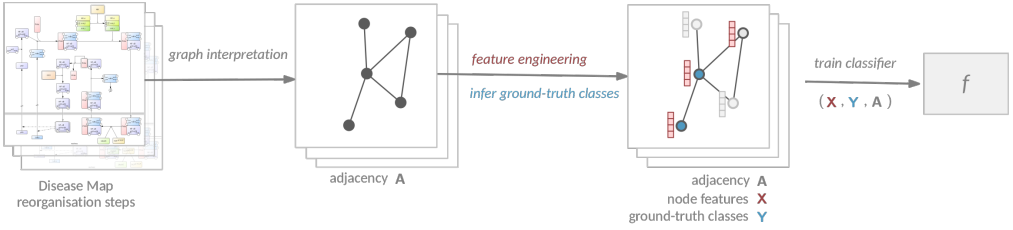
ReconMapOlder (collapsed)

4030 species aliases, 5535 reactions
3089 (76.65%) for prediction
741 (23.99%) to-be-duplicated

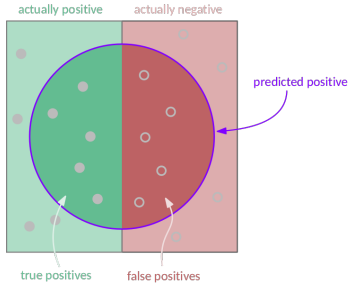


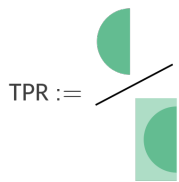


“Supervised node classification”?



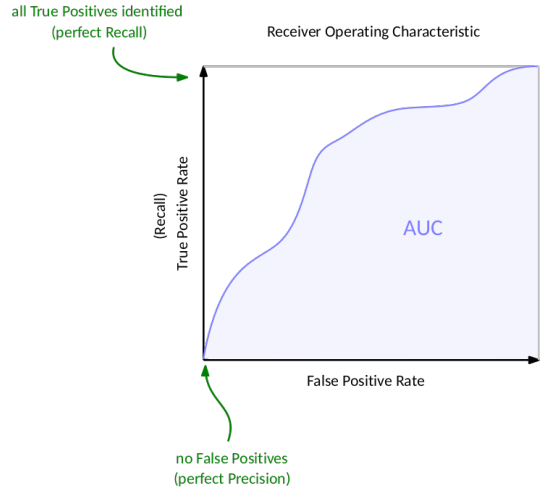
- To compare classifiers, we need an **unbiased performance measure**
- Classifiers used herein yield a **confidence score** in $[0, 1]$ for a given example
- Obtain concrete classification by setting a **decision threshold**, yields
  - **True Positive Rate** (TPR): # true positives/# actually positive
  - **False Positive Rate** (FPR): # false positives/# actually negative





- Focus on positive class
- Insensitive to class imbalance

- 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
- Usually a tradeoff, choice depends on use-case
  - Accept only few high-confidence predictions $\rightarrow$ low FPR, but also low TPR (Recall)
  - Lower decision threshold $\rightarrow$ increase TPR at cost of increased FPR



Plot TPR and FPR as function of decision threshold

maybe recap slide here (then: ok, so now we have all the background we need to actually do stuff)

Previous work by **?** [**?**]:

- Node features based on graph centralities
- Consider collapsed map plus reorganisation steps
- Supplied to Support Vector Machine classifier

We extend this in several directions:

- Explore different classifier (Graph Neural Networks)
- Explore importance of reorganisation steps
- Explore choice of features
- Heuristic for determining number of duplicates and edge attachment

(data setup: train on ADReorg (collapsed plus reorg), evaluate on both PDMap and ReconMap – because interesting to see how they behave differently) (note on how we always build on the previous experiment)

reproducing: show AUC comparison for PDMap, ROC curves for ReconMap

GNN recap & motivation of GNN

additionally show curves for GNN model (emphasize that this is naive approach)

discussion and takeaway of first experiment

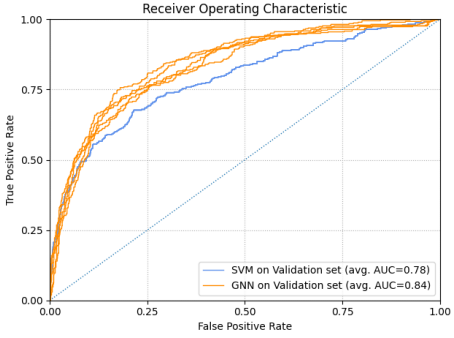
importance of reorganisation steps

importance of message-passing

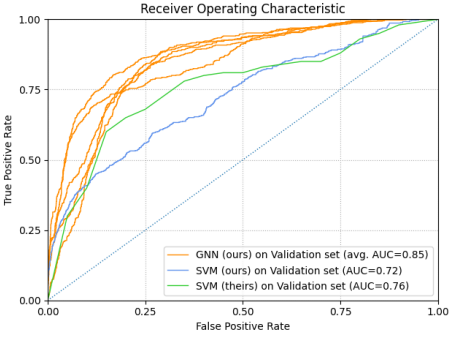
feature selection

GO features

attachment



(a) (AlzPathwayReorg → PDMap)



(b) (AlzPathwayReorg → ReconMap)

foo

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