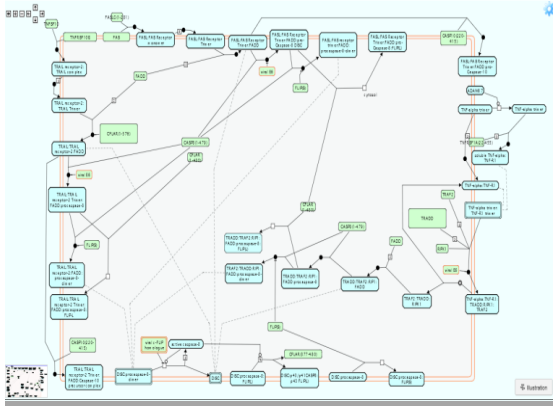
Using ILP to Identify Pathway Activation Patterns in Systems Biology

samuel.neaves@kcl.ac.uk

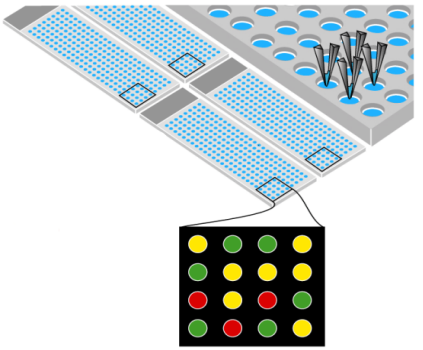
Aims

- To describe the system biology differences between one disease state or another.
- In this case two types of Lung Cancer.

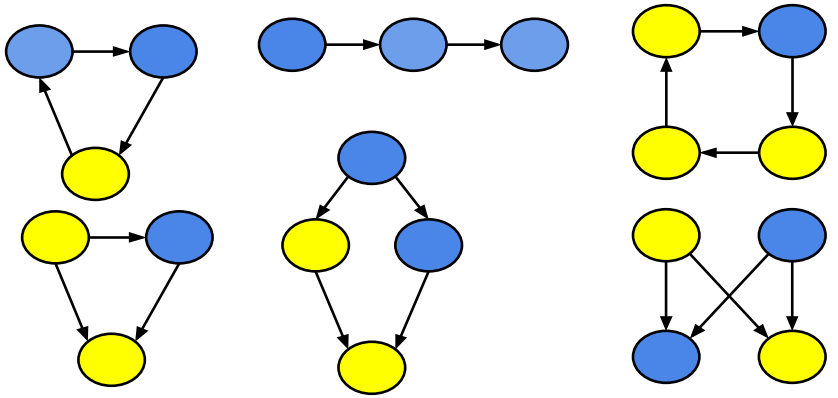
Background pathway data



Miroarray Transcriptome Data

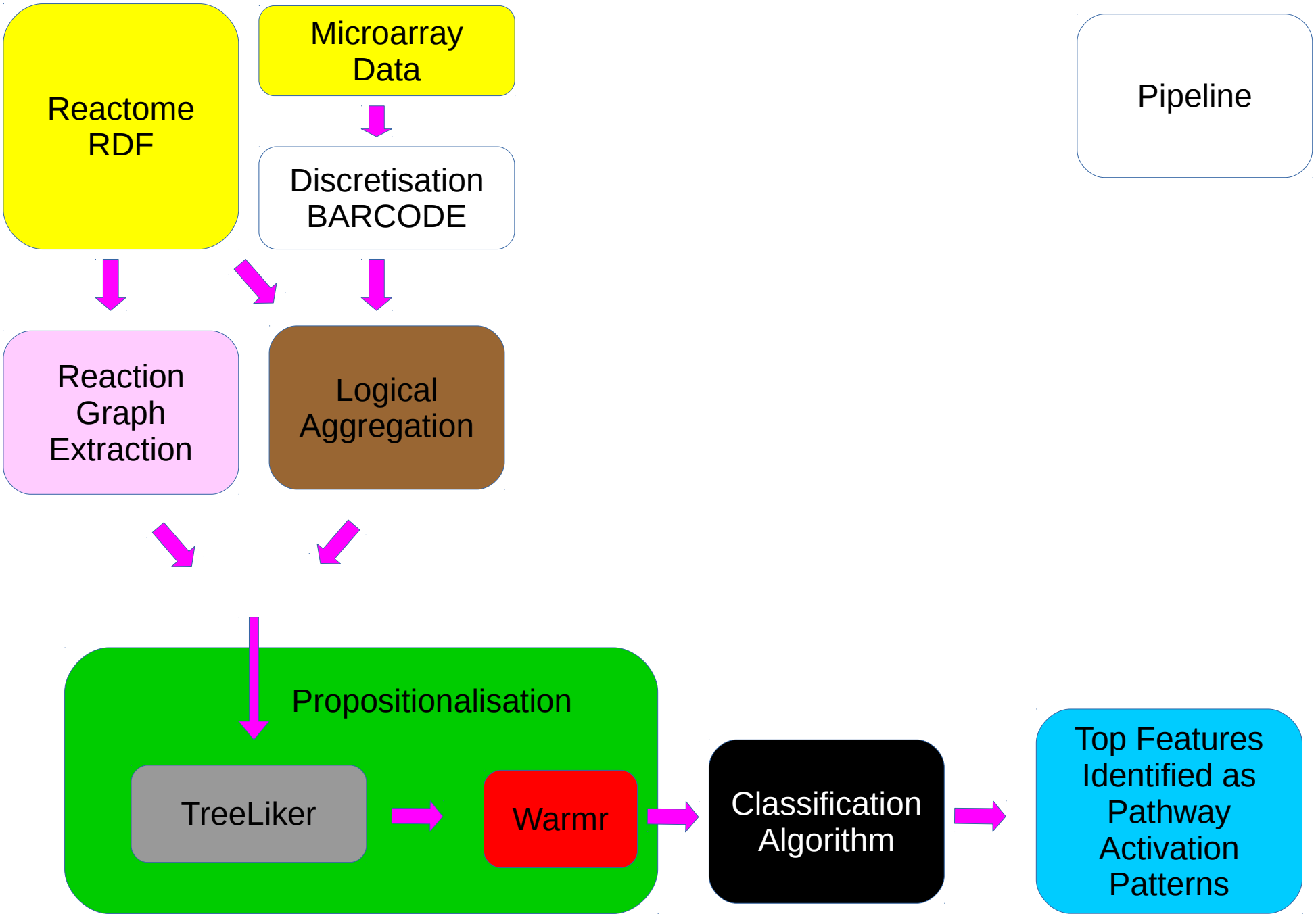


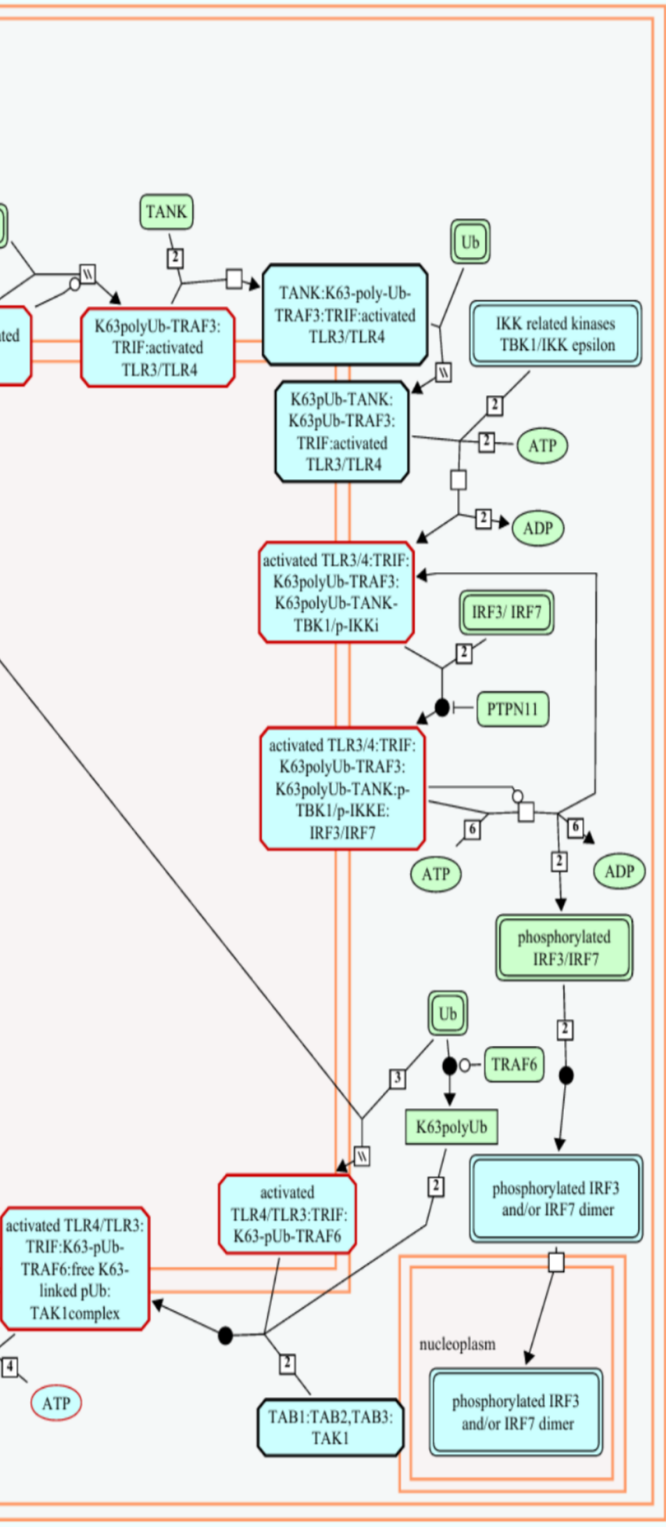
Pathway Activation Patterns

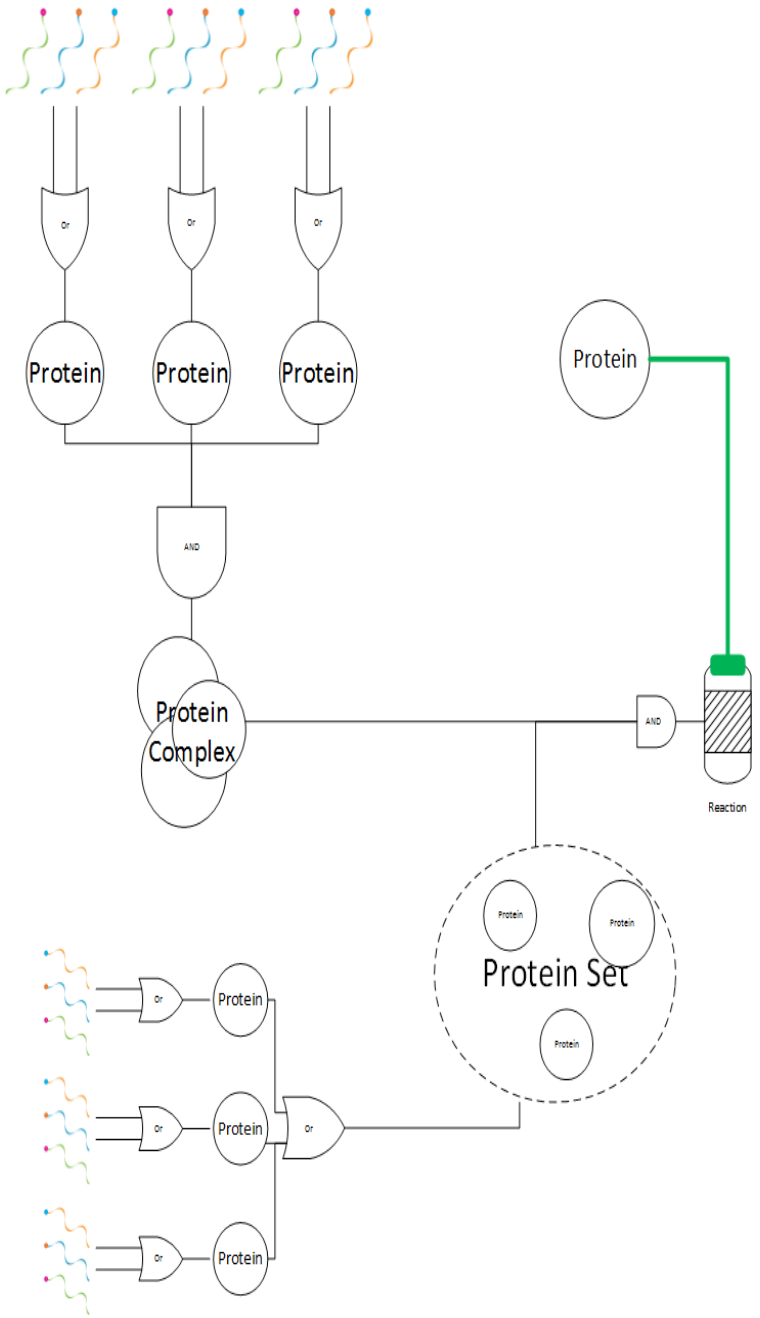


Aims

- We are not aiming to build a diagnostic classifier.
- For example in Lung cancer we would know if a patient has one type of cancer or another by the nature of the biopsy.
- We want instead to understand what the system level perturbations are between the cases.
- We want models that are understandable and provide biological insight.
- For example, longer more complicated patterns are of greater interest to a biologist than simple short patterns.

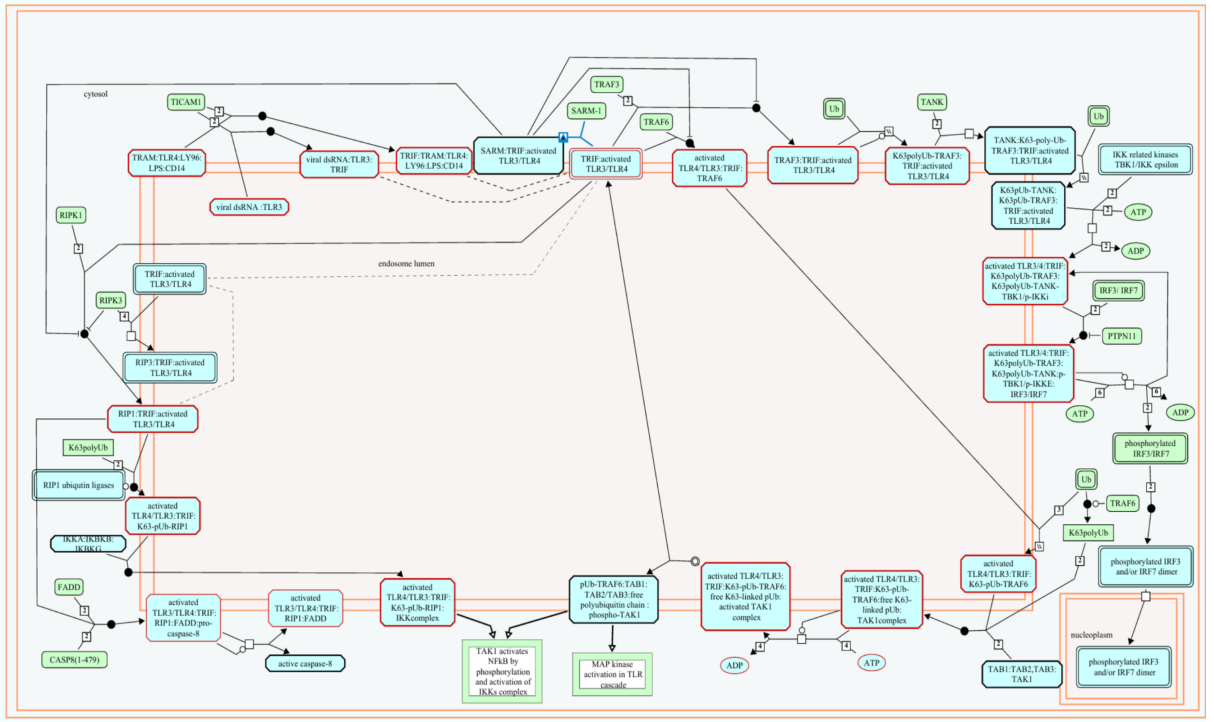

Reactome RDF
Microarray Data
Discretisation BARCODE
Pipeline
Reaction Graph Extraction
Logical Aggregation
Propositionalisation
TreeLiker
Warmr
Classification Algorithm
Top Features Identified as Pathway Activation Patterns



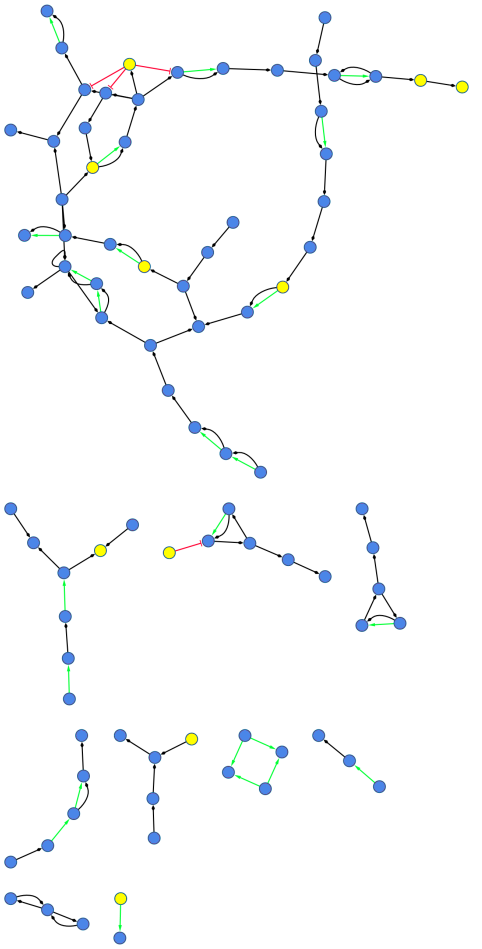


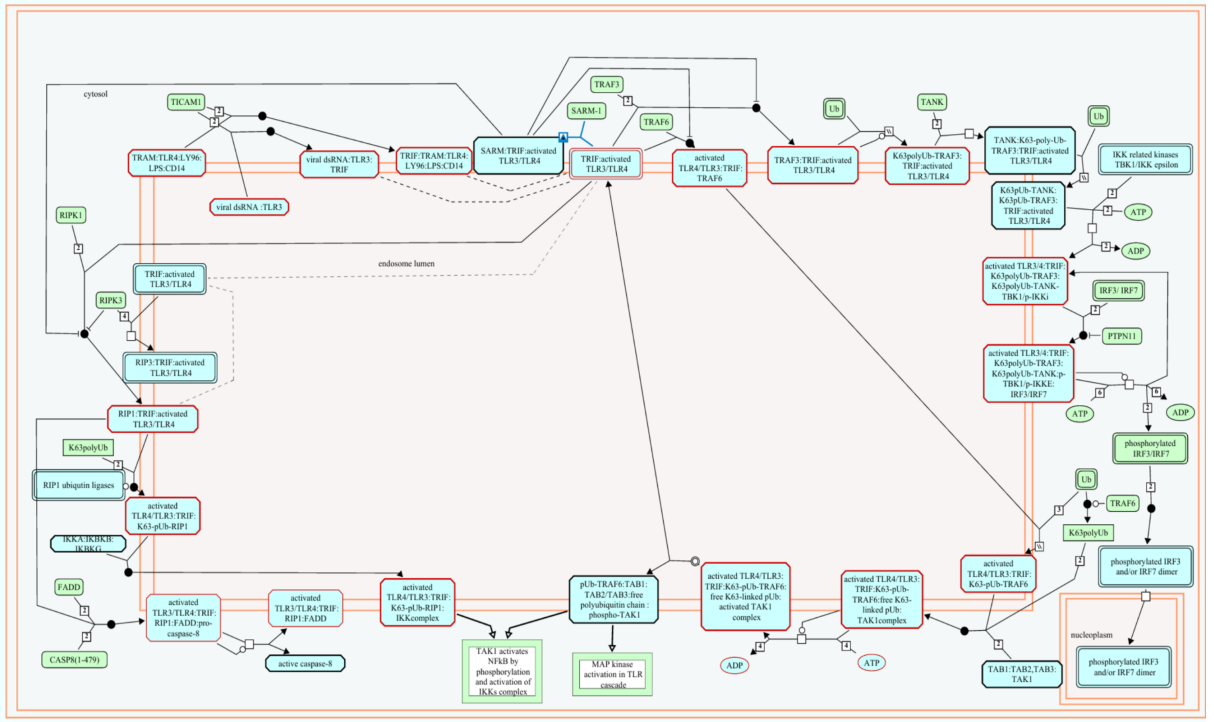
Logical Aggregation

We say a reaction'*is active*' if all its inputs are ‘on’ and all its activation controls are ‘on’ and none of its inhibition controls are ‘on’

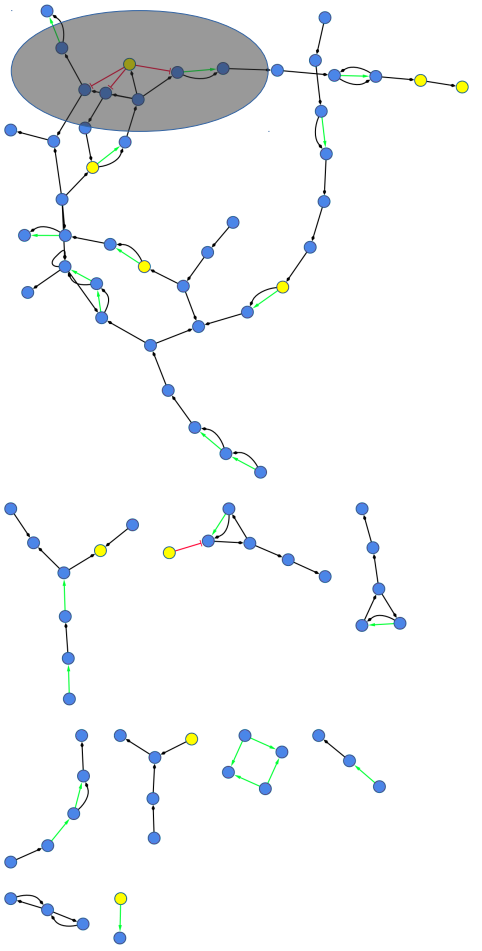


Reaction Graph Extraction

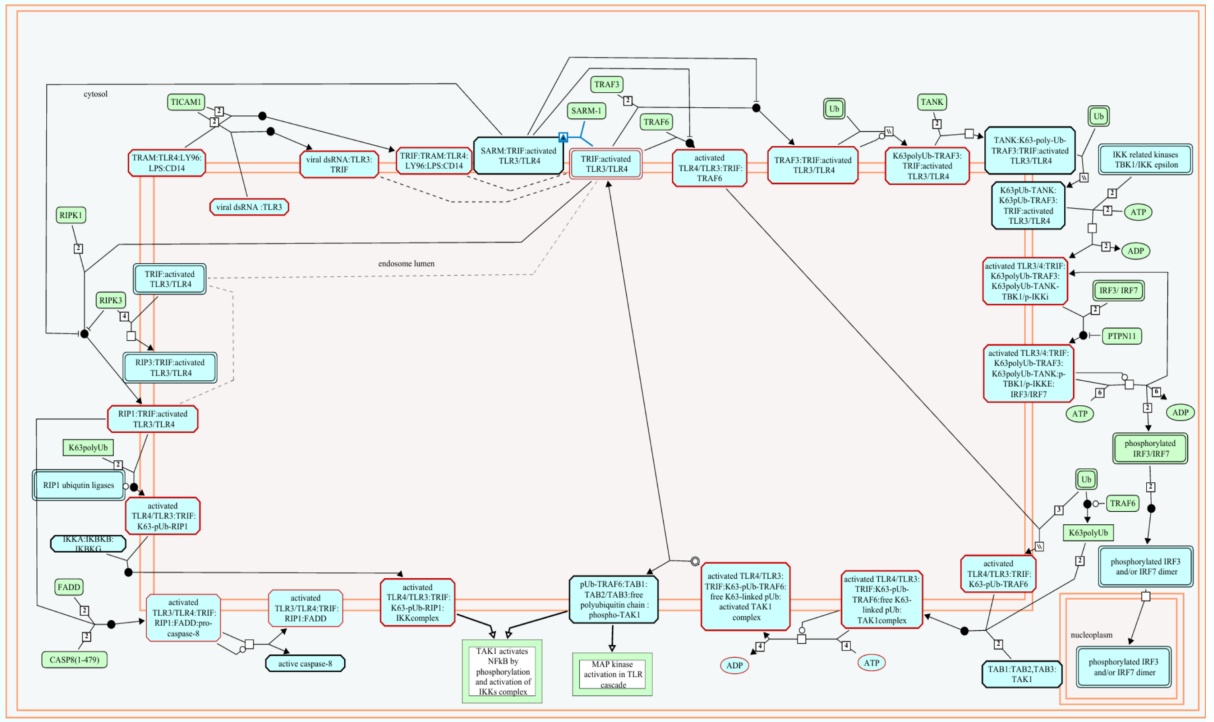


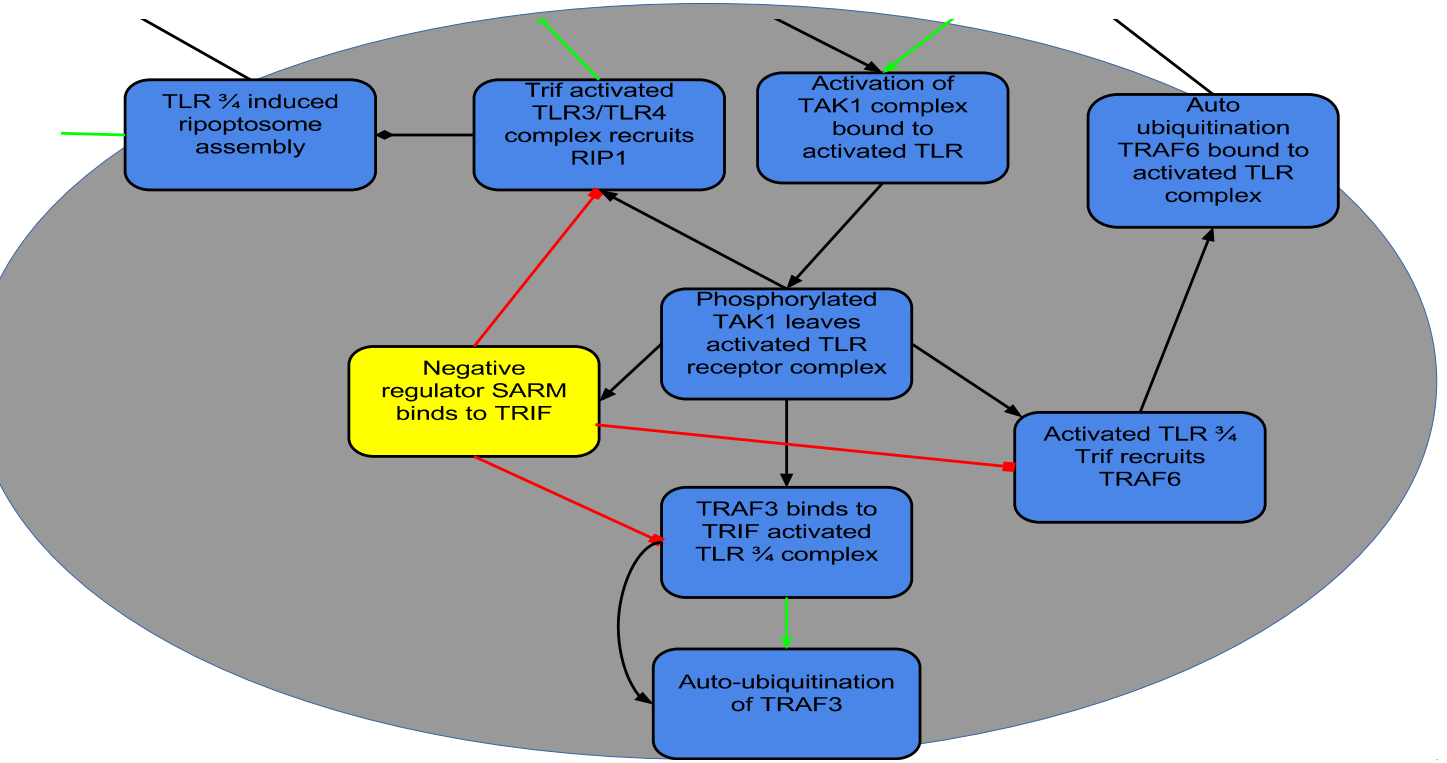


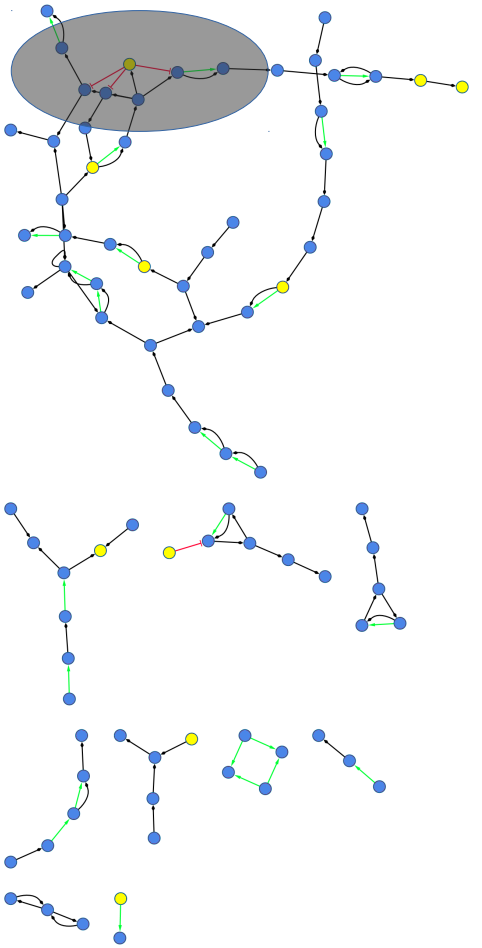
Reaction Graph Extraction



Reaction Graph Extraction

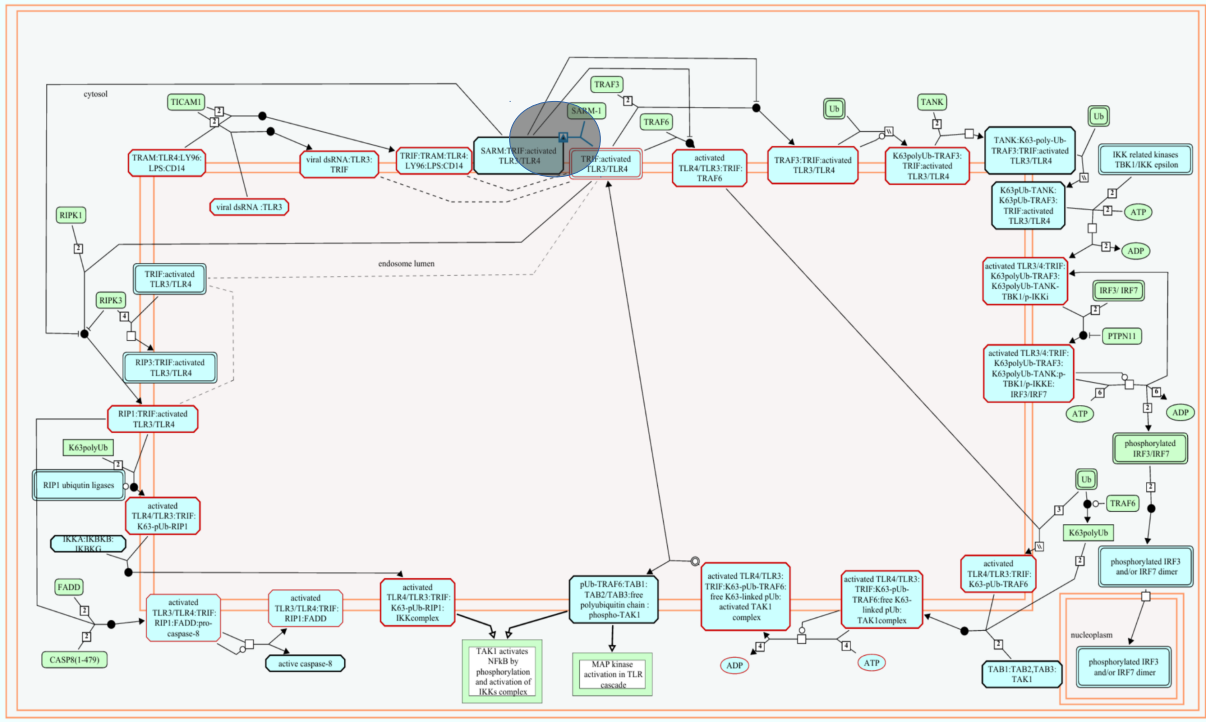


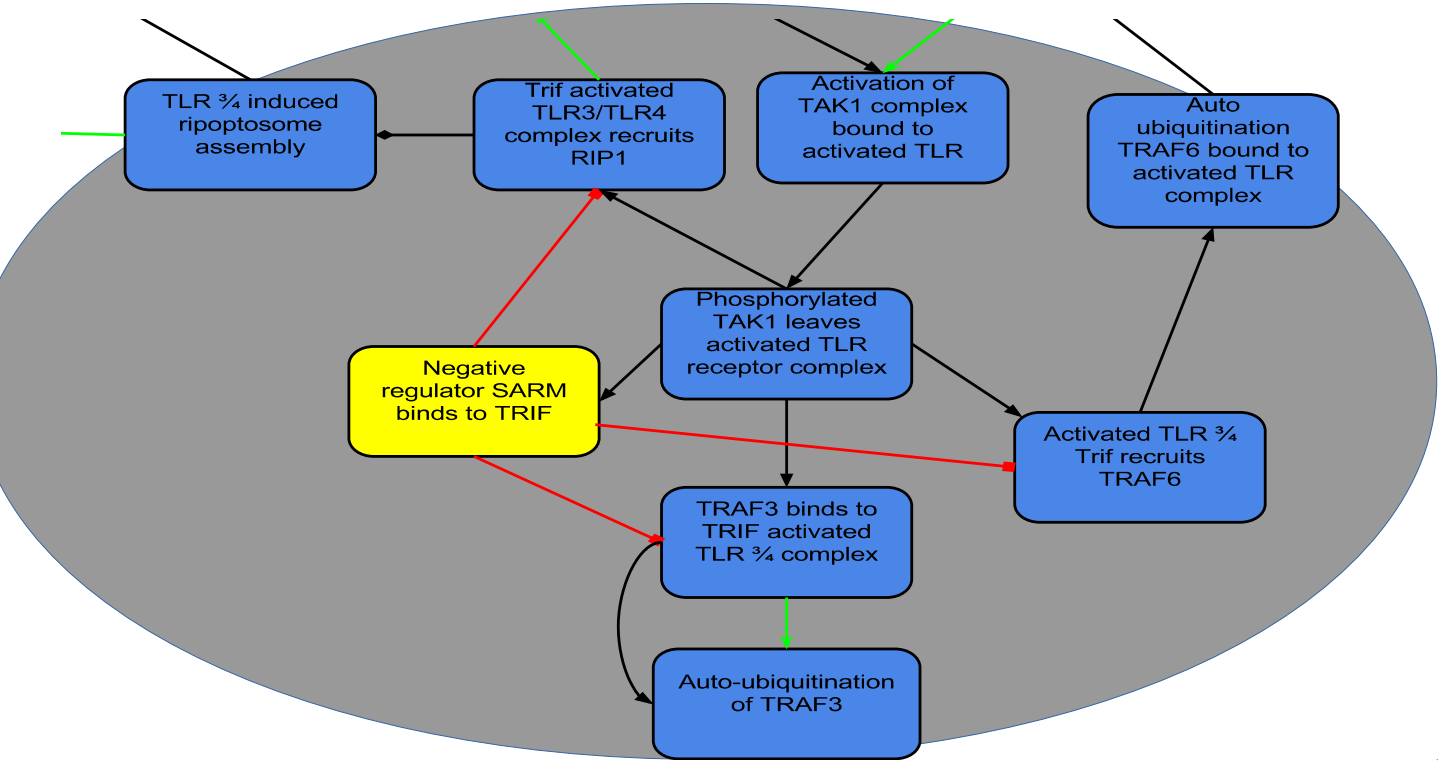


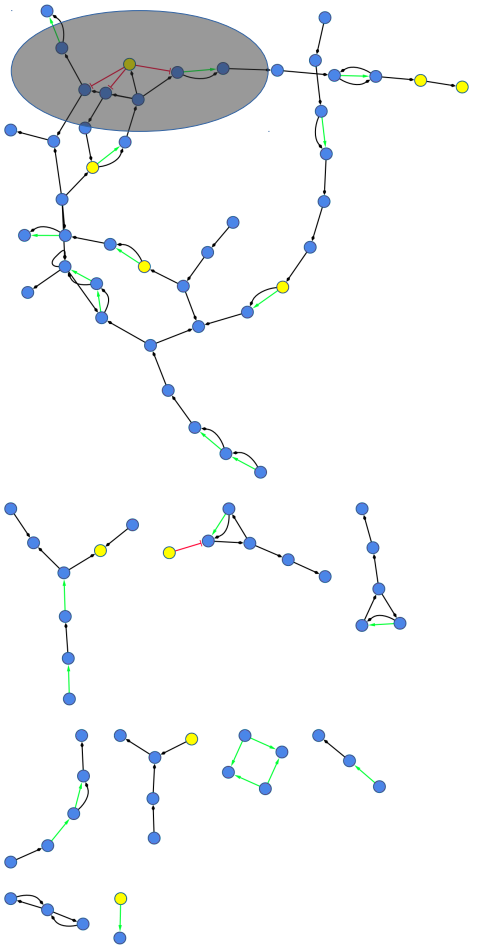


Reaction Graph Extraction

Each example is now a boolean reaction network.







Propositionlisation

Propositionlisation methods are similar to searching for discriminative SQL queries.

Each query is either true or false for a given example.

Propositionlisation transforms each example from the boolean network into a vector of binary values where each feature corresponds to a discovered query.

From this classical attribute value learners can be used.

We have experimented with two tools Warmr and Treeliker, and a combined method.

Warmr		TreeLiker (RELF algorithm)
Level wise search for frequent queries.	⟷	Finds features by combining small conjunctive blocks.
Pruning is not done by relevance, so many irrelevant queries found.	⟷	Able to find long discriminative features, but features are limited in that they can not contain cycles.
Powerful control of the language bias. Allows full Prolog programs as background knowledge.	⟷	It is harder to use background knowledge and to control the language bias.

Combined approach

Propositionlisation

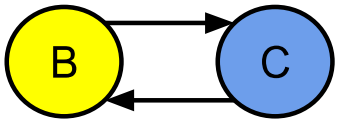
Initially use Treeliker to find long discriminative treelike features.

Next use the language bias facilities of Warmr to start a search beginning with a top treelike feature.

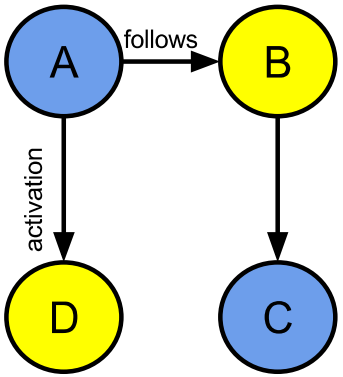
Add further language bias to Warmr to guide the search to join up cycles where possible.

Extending Treeliker found features

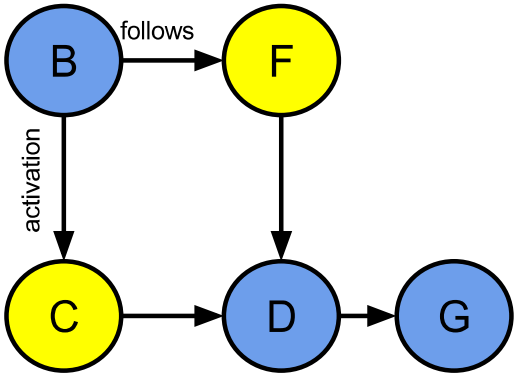
Propositionlisation



Feature 1



Feature 2

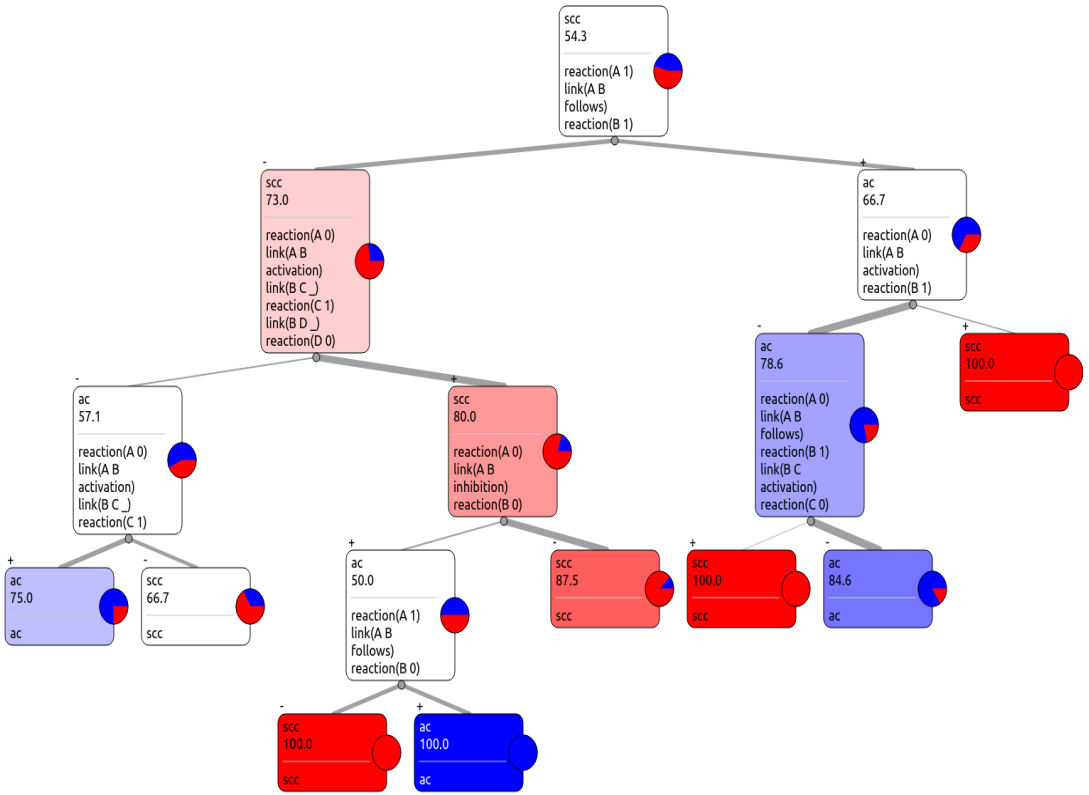


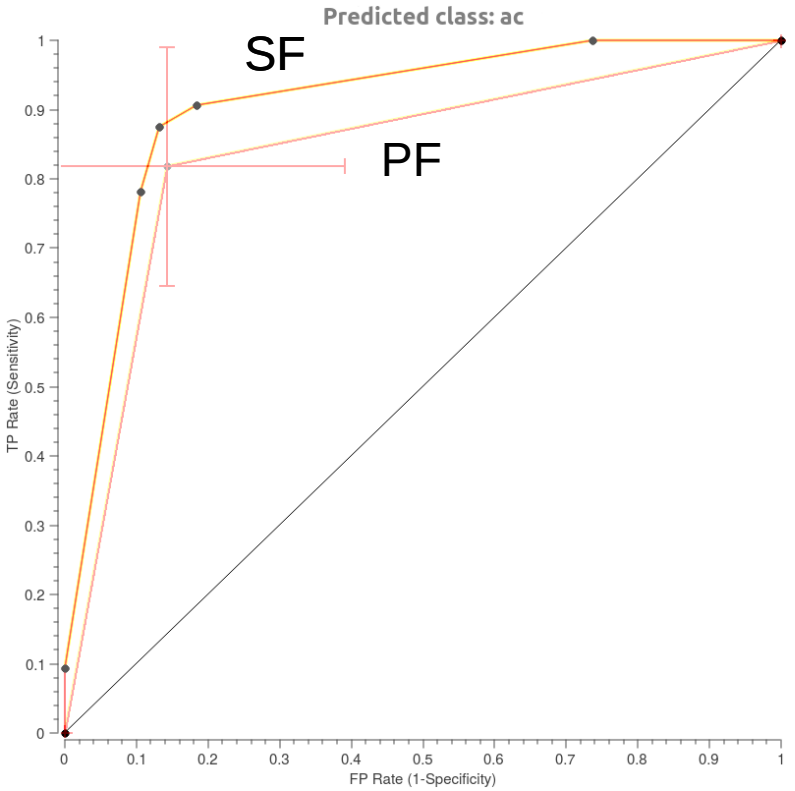
Feature 3

Feature 1 found by Warmr contains a cycle, but is short.
Feature 2 found by Treeliker, is a longer feature but does not contain a cycle.
Feature 3 found by combined method is both long and contains a cycle.

Classification Algorithm

Top Features Identified as Pathway Activation Patterns





SF= Structured Features
PF= Raw Probe Features

Summary

- We have shown how ILP methods can find Pathway Activation Patterns.
- We did this by using Reactome as background knowledge.
- We used Logical Aggregation to Reactions.
- We used Propositionalisation to find structural features.
- We identified top structural features from a classification model as Pathway Activation Patterns

Questions?

- Contact me: samuel.neaves@kcl.ac.uk

Number of features

- Probe table approximately 55k features
- Reaction table approximately 6.5k features
- One Pathway (approx 1k pathways) has 28 Treelike features. (Depends on language bias)
- Combined approach will make a number of features from each Treelike feature. (Depends on Treelike feature and language bias).