Predicting yeast properties from knowledge graphs using GNNs and box embeddings

Filip Kronström[1], Daniel Brunnsåker[1], Alexander H. Gower[1], Ievgeniia A. Tiukova[12] and Ross D. King[134]

[1] Chalmers University of Technology, Gothenburg, Sweden
[2] KTH Royal University of Technology, Stockholm, Sweden
[3] Cambridge University, Cambridge, United Kingdom
[4] Alan Turing Institute, London, United Kingdom

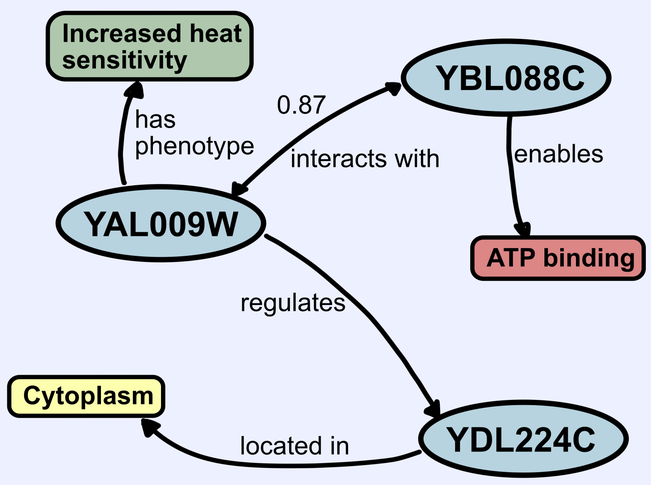
Overview

We construct a Knowledge Graph (KG) about the yeast *Saccharomyces cerevisiae* combining information from various sources. Based on ontologies defining class hierarchies for the terms used box embeddings are found for the nodes in the graph. We demonstrate that Graph Neural Networks (GNNs), combining the KG and box embeddings can be used to predict the effect of gene deletions in *S. cerevisiae*. Insights from these models can drive scientific discovery, for example through the embeddings generated.

Knowledge Graph

A Knowledge Graph (KG) is created using data from Saccharomyces Genome Database, BioCyc, and The Cell Map. This KG contains information about genes and reactions in the yeast *S. cerevisiae* expressed in controlled vocabularies.

Interaction relations between genes are accompanied by a real number describing the growth when these genes are knocked out [1].

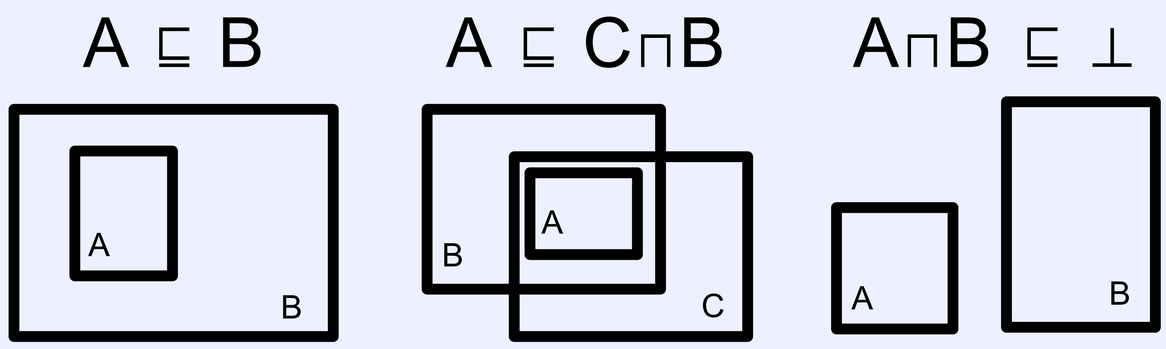


Box embeddings

The terms in the KG are expressed as classes from various ontologies defining class hierarchies. We represent the nodes (classes) in our KG as axis-aligned hyperrectangles which are found by optimizing the conditional probabilities for class inclusions [2]. The classes in our KG can be divided into nine distinct class types for which separate box embeddings are found.

A class inclusion axiom, $A \sqsubseteq B$, corresponds to $P(B|A) = 1$ which is calculated as

$$P(B|A) = \frac{\mathrm{Vol}(\mathrm{Box}_A \cap \mathrm{Box}_B)}{\mathrm{Vol}(\mathrm{Box}_A)}$$

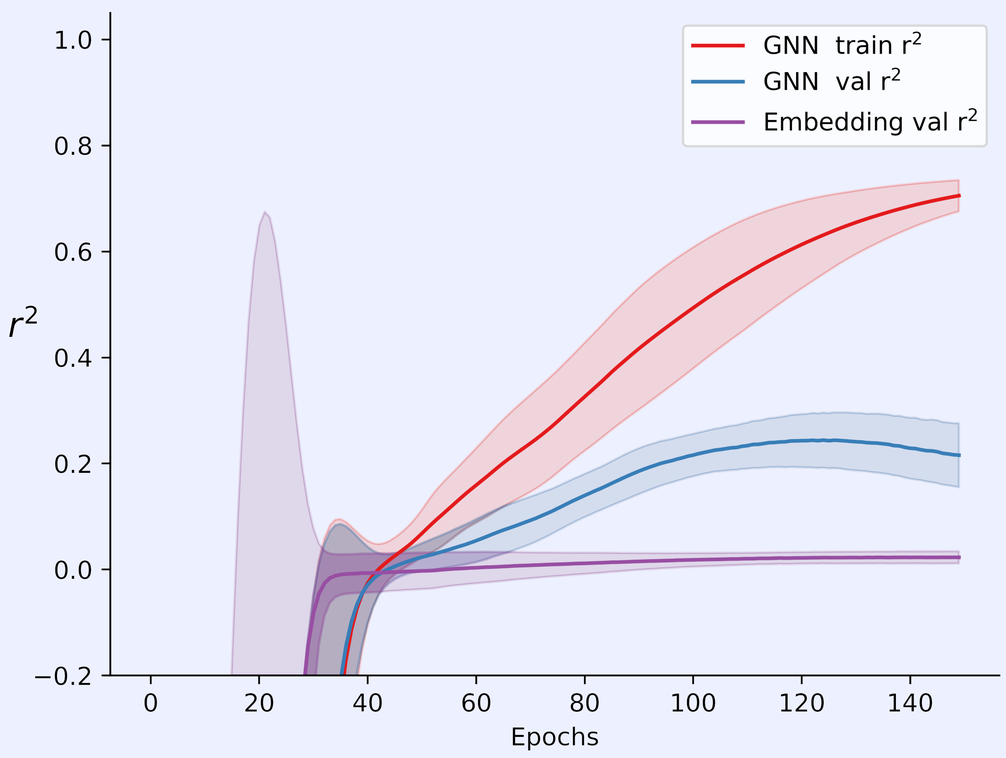


Graph Neural Network

Heterogeneous GNNs are used to propagate relation information in the KG predict properties given the node embeddings and edges in the KG.

During the GNN training class box embeddings are altered, but penalised when not adhering to the class hierarchies from the ontologies.

These models can be used for, e.g., gene function prediction (link prediction) or, as shown below, growth prediction for double gene knock-outs.

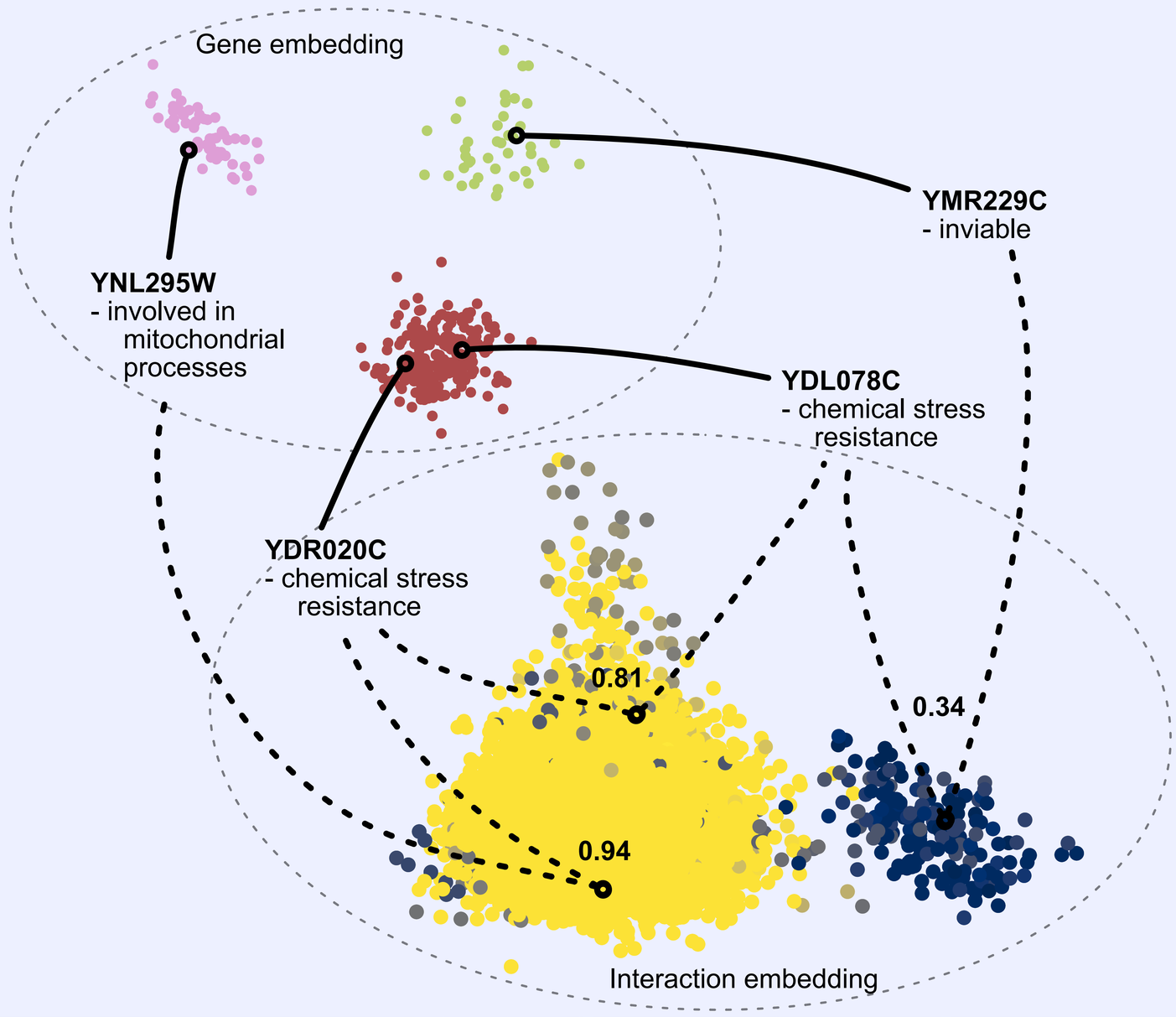


r² values for 10-fold cross validation of growth when knocking out pairs of genes. All genes in the validation set are unseen during training. The predictions for the comparison 'Embedding' model are based only on the box embeddings of the genes.

Scientific discovery

Embeddings generated by the GNNs combine information from box representations and graph neighbours. Below clusters in the embeddings of genes and their interactions (the Hadamard product between two genes) are shown.

Embeddings can be useful tools for scientific discovery as they for instance allow for comparison of new knowledge to the facts asserted in the KG.



[1] Costanzo, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science 353.

[2] Dasgupta, et al. 2020. Improving local identifiability in probabilistic box embeddings. NeurIPS '20.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM