BERTwalk for integrating gene networks to predict gene- to pathway-level properties

Rami Nasser, Roded Sharan

Agenda

- Problem Formulation
- Previous Work
- BERTwalk
- Results

Notations

- $G = (V, E, w)$ a graph
  - V set of nodes
  - E set of edges
  - $w : E \to \mathbb{R}$ weights of edges
- A adj matrix: $A_{ij} = w(i, j)$ for $(i, j) \in E$, else $A_{ij} = 0$.
- $D_{ii} = \sum_j A_{ij}$ the diagonal degree matrix.
- $X \in \mathbb{R}^{|V| \times e}$ feature matrix for nodes.
- Network propagation (RWR): $X' = (1 - \alpha)WX + \alpha X$
  - $W = D^{-1/2} A D^{-1/2}$

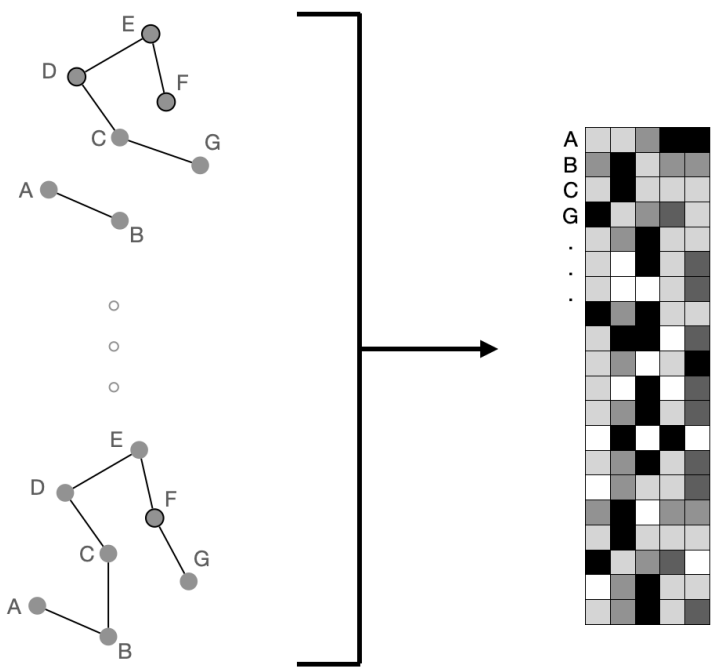
Problem Formulation

Given

A collection of T networks $\{G_1, G_2, \cdots G_T\}$ on the same set of n nodes, and a number e of desired embedding dimension.

Goal

learn a node embedding matrix $X \in \mathbb{R}^{n \times e}$



Evaluation Criteria - BIONIC

Tasks

- Module detection: hierarchically clustered.
- Gene function prediction.

Annotated Collections of Proteins

- IntAct
- KEGG
- GO

Data - 3 Yeast Networks

Costanzo et al

A network of correlated genetic interaction profiles.

Hu et al

A co-expression network.

Krogan et al

A protein-protein interaction network.

	Costanzo	Hu	Krogan
# nodes	4,529	1,101	2,674
# edges	33,056	14,826	7,075

Previous Work and new challenges

Popular Methods

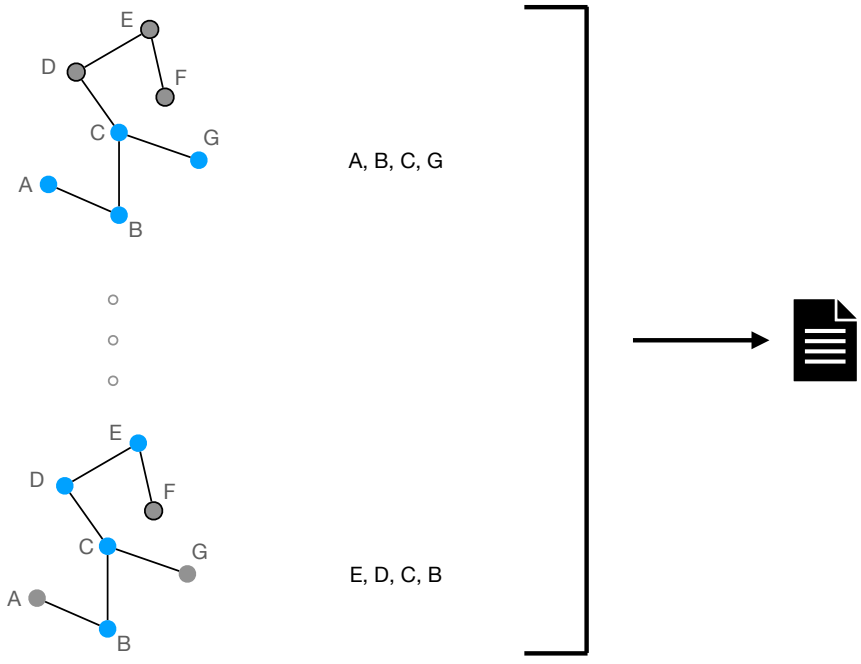
- Auto-Encoder methods. (BIONIC)
- Spectral methods.
- Matrix Factorization.
- Node2vec.

Limitations & Challenges For Networks integration

- GNN limited to number of distant nodes.
- The final embedding is averaged over the networks.
- Decoding is defined based on inner product.

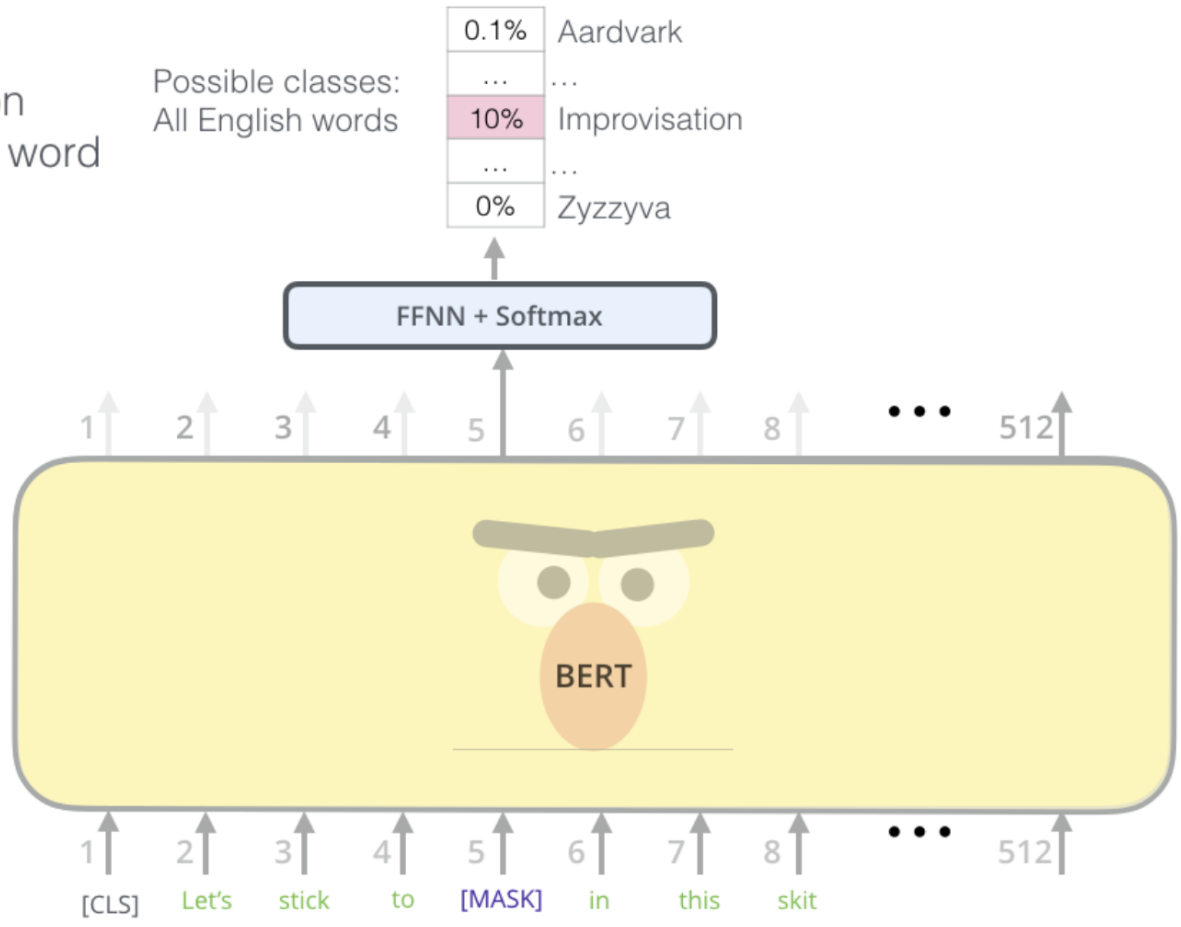
From Graph to Text

- DeepWalk, node2vec.
- 5 walks from each node in every network.
- Length of walks 10.



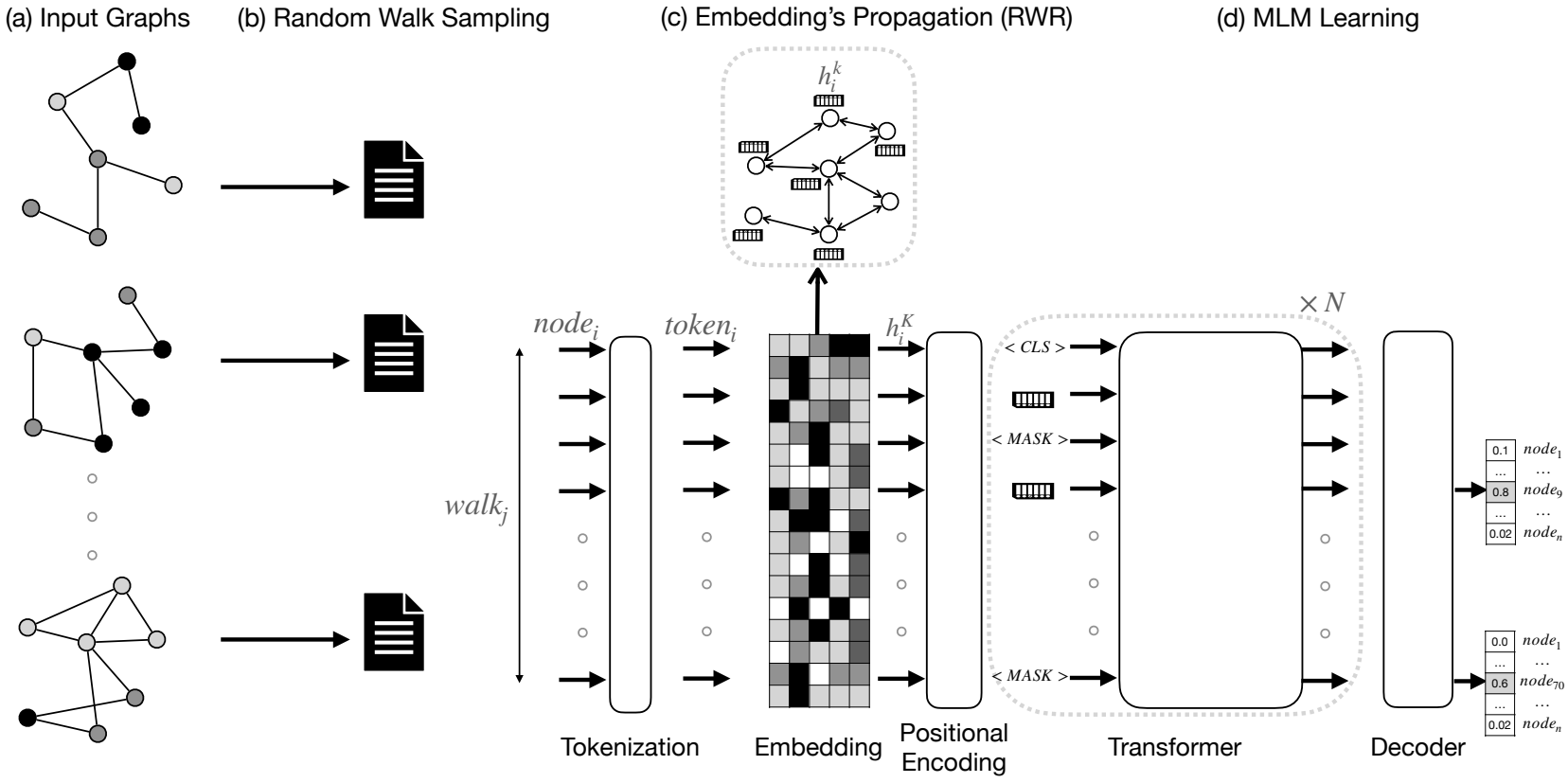
Masked Language Modeling

Use the output of the masked word's position to predict the masked word

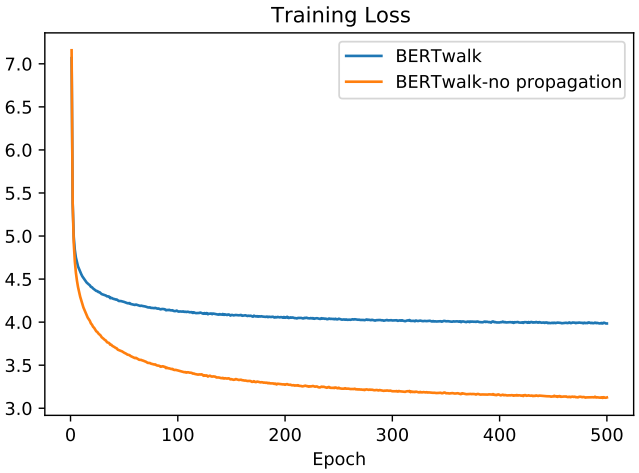


https://jalammar.github.io/illustrated-bert/

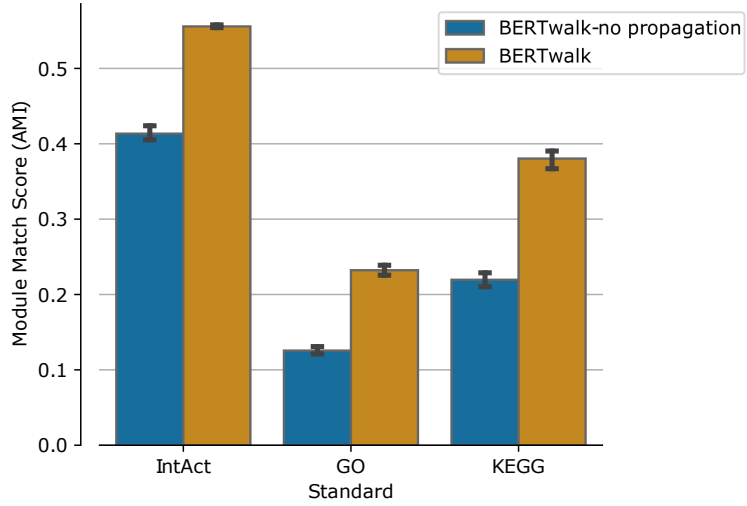
BERTwalk


(a) Input Graphs
(b) Random Walk Sampling
(c) Embedding's Propagation (RWR)
(d) MLM Learning
h_i^k
$node_i$
$token_i$
h_i^K
$walk_j$
$\times N$
< CLS >
< MASK >
< MASK >
0.1
0.8
0.02
$node_1$
$node_9$
$node_n$
0.0
0.6
0.02
$node_1$
$node_{70}$
$node_n$
Tokenization
Embedding
Positional Encoding
Transformer
Decoder

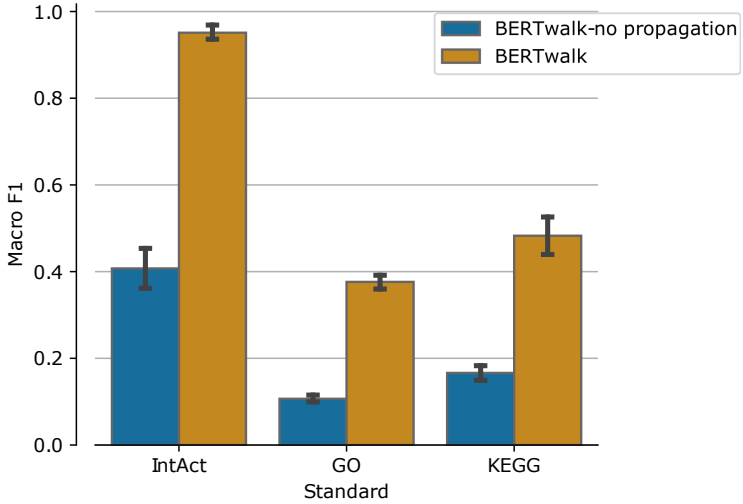
Effect of Propagation During Training



(a) CrossEntropy Loss

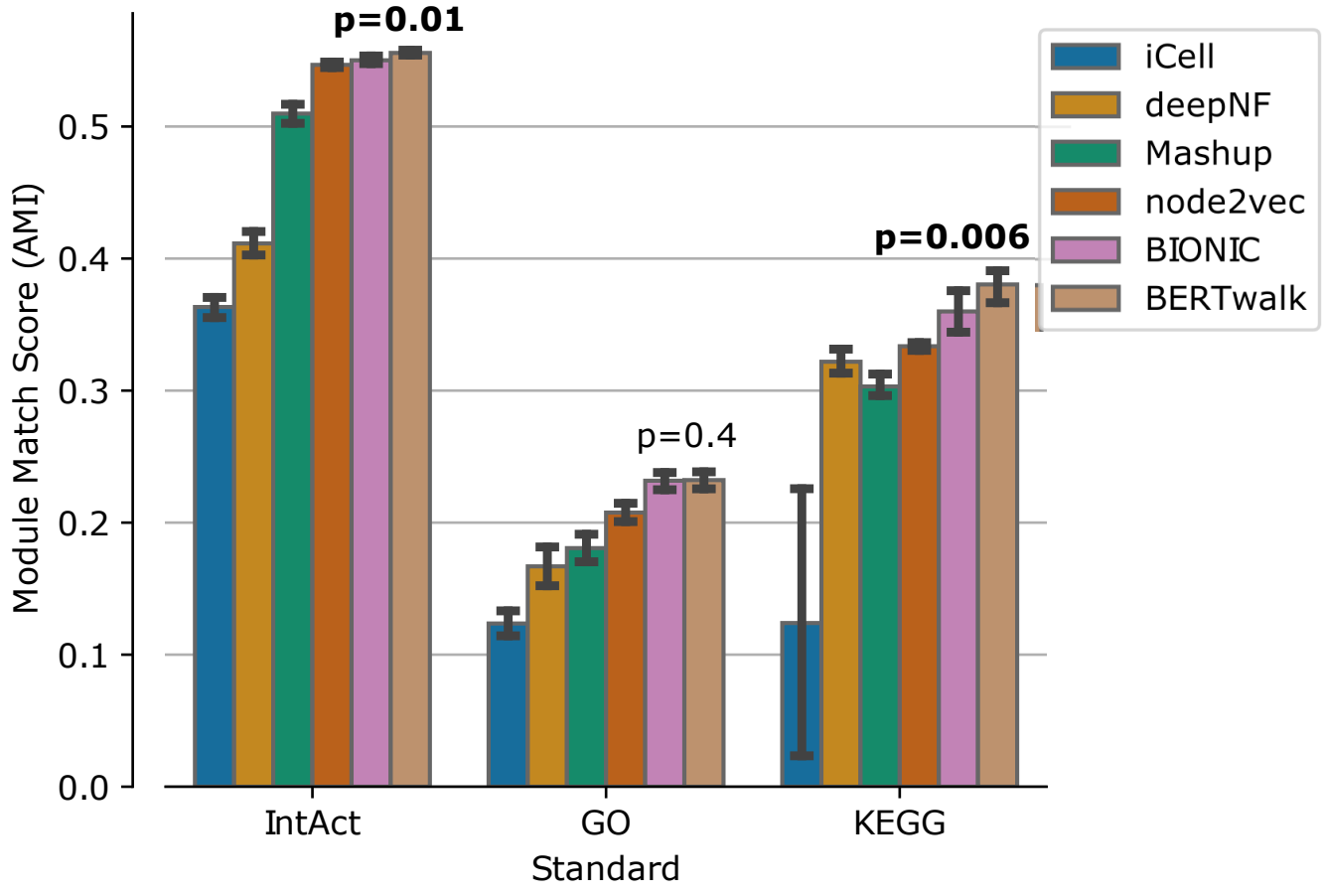


(b) Module detection

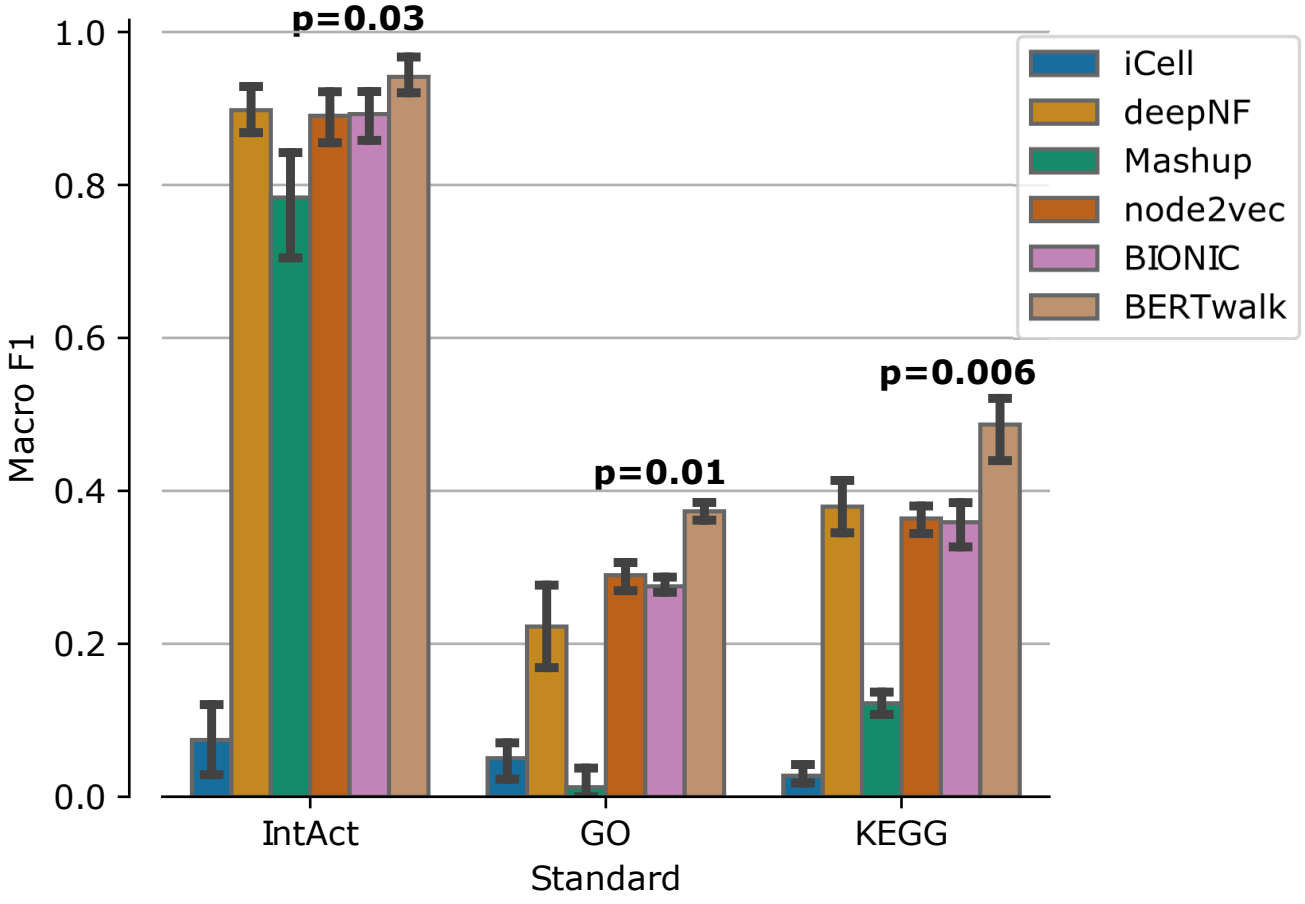


(c) Function Prediction

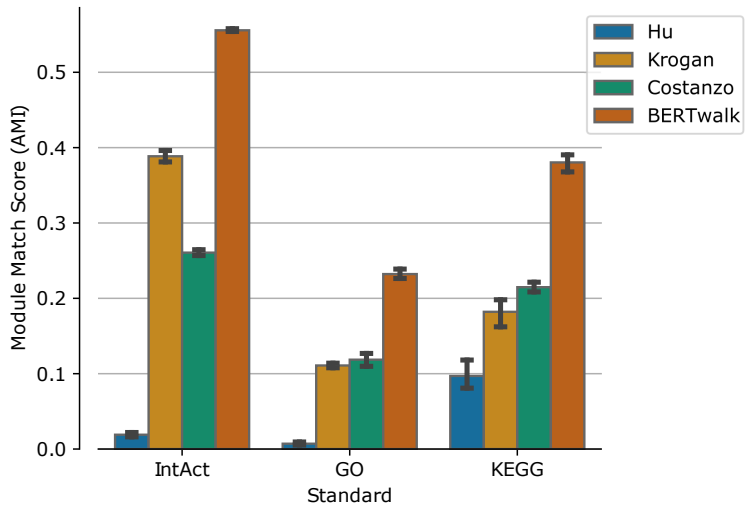
Results - Module Detection



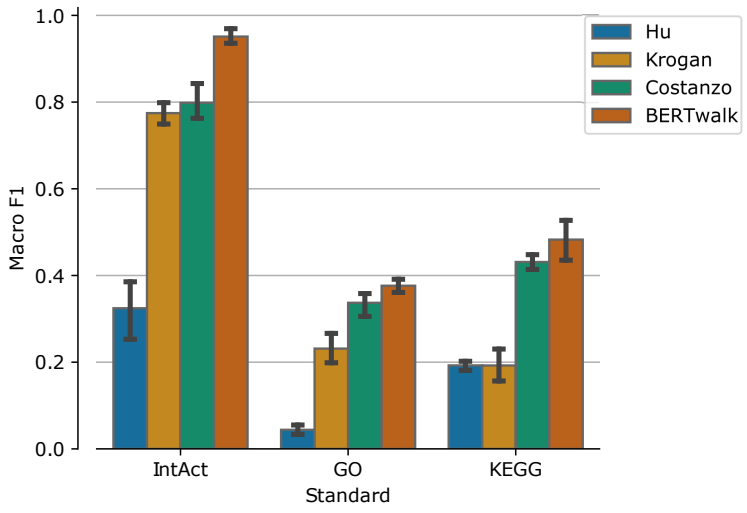
Results - Gene Function



Results - Power of Integration



(a) Module detection



(b) Function prediction

Pathway Prediction

- Synthesized data using PPI (Krogan) and knockout gene expression (Holestege) data.
- Paths that start with a deleted mutant and end in a deferentially expressed gene are labeled as positive and the rest are labeled negative
- In total we constructed 26740 paths.
- 18% of paths are positive.
- In BERTwalk, using [CLS] token.
- In BIONIC, averaged the embedding of the path's nodes.

Results - Pathway Prediction