EGRC: Efficient inferring Gene Regulatory Network based on graph convolution network

Duaa Mohammad Alawad, Ataur Katebi[2,3] , Md Tamjidul Hoque

email: dmalawad@uno.edu, a.katebi@neu.edu, thoque@uno.edu

[1]Department of Computer Science, University of New Orleans, New Orleans, LA, USA.

[2]Department of Bioengineering, Northeastern University, Boston, MA, USA.

[3]Center for Theoretical Biological Physics, Northeastern University, Boston, MA, USA.

Introduction

- Different cell types have distinct gene expression profiles, and cells differentiate from one cell state to another by modifying their expression profiles via gene transcription.
- Transcription factors (TFs) are protein molecules that regulate the transcription of a multitude of target genes.
- A network of TFs and their target genes are responsible for the normal function of a biological process.
- These TFs and their target genes form unique patterns.
- We developed a graph neural network-based method, named EGRC, that learns these patterns from TF-target relationships and then predicts TF regulatory networks from gene expression data.

Materials

Table 1: DREAM4 and DREAM5 Datasets.

Dataset	No. of TFs	No. of Genes
DREAM5 Network (*In silico*)	195	1643
DREAM5 Network (*S. cerevisiae*)	333	5949
DREAM5 Network (*E. coli*)	334	4511
DREAM5 Network (*In silico*)	195	1643
Network 5	30	392

Performance evaluation metrics

Table 2: Name and definition of Performance Evaluation Metrics

Name of Metric	Definition
Area under curve (AUC)	Area under the receiver operating characteristic curve
AUPR	Area under the precision-recall curve

Conclusion

- We developed a framework named EGRC to distinguish whether the extracted subgraph centered at two nodes contains the link between the two nodes.
- A pair that links between a transcription factor (TF) and a target gene and their neighbors are labeled as a positive subgraph, while an unlinked TF and target gene pair and their neighbors are labeled as a negative subgraph.
- According to the experiment's findings, the following factors may contribute to the GRN prediction power of the EGRC:

  (1) using an ensemble of heuristic skeletons.

  (2) using graph embedding can capture the topological structures of the network to predict the links.

  (3) improving the pooling layer (SAGPool technique) can improve a graph classifier's performance.
- We believe that the ability of our proposed method to infer GRN with higher accuracy will have a more significant impact on understanding biological systems and disease processes.

Methods

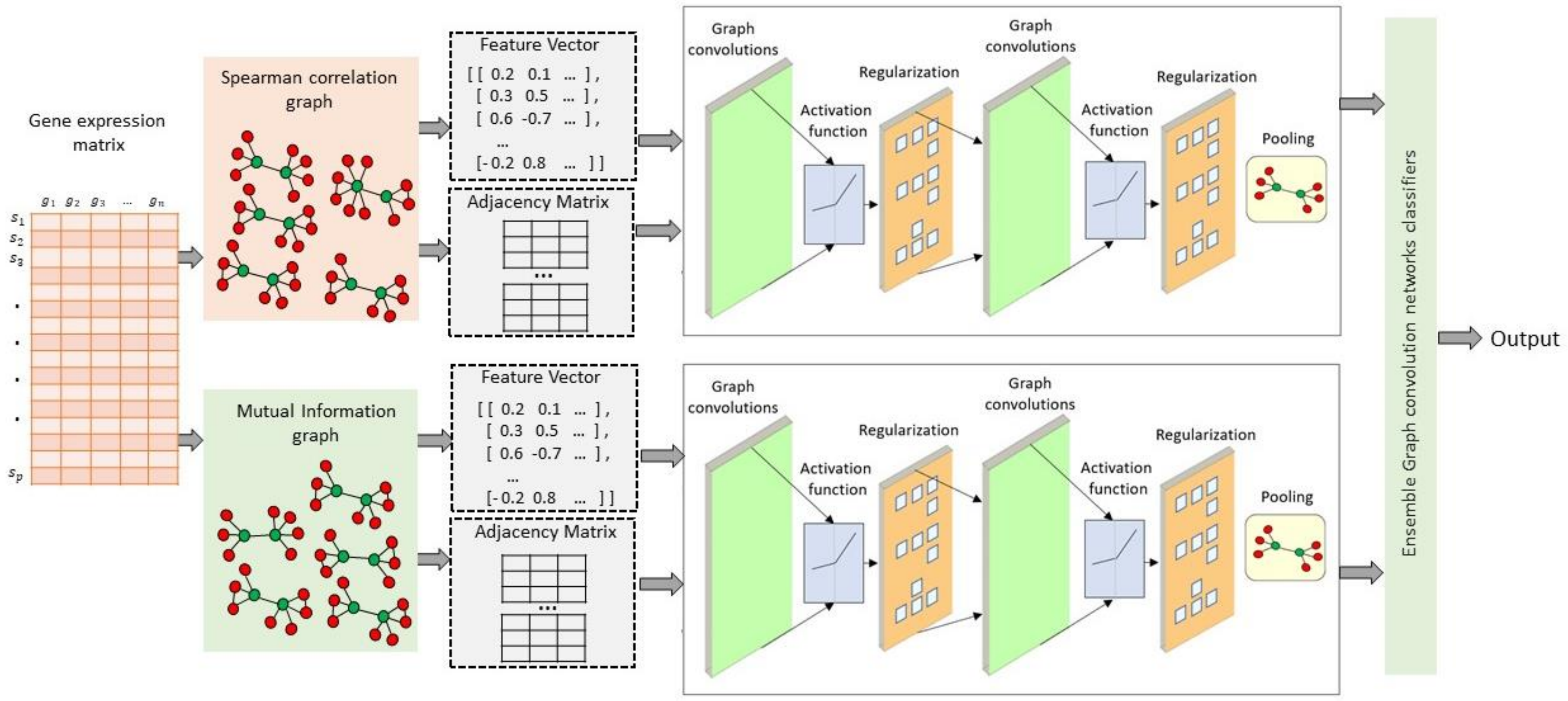


Figure 1:EGRC Framework. EGRC scheme. Noisy starting skeletons are derived from Spearman's correlation and mutual information

- The four phases make up the entire EGRC process:

1) Constructing noisy skeletons

- **Heuristic approaches** [such as **Spearman's correlation** and **mutual information**] are used to infer associations between TFs and their target genes using the gene expression.
- Spearman correlation and Mutual information can be used to compute the **nonlinear correlation between two genes**.
- Many **bipartite graphs** $< TF, G, L >$
  - The thresholds are used in :
    - Spearman's correlation = 0.8
    - Mutual information = 0.5

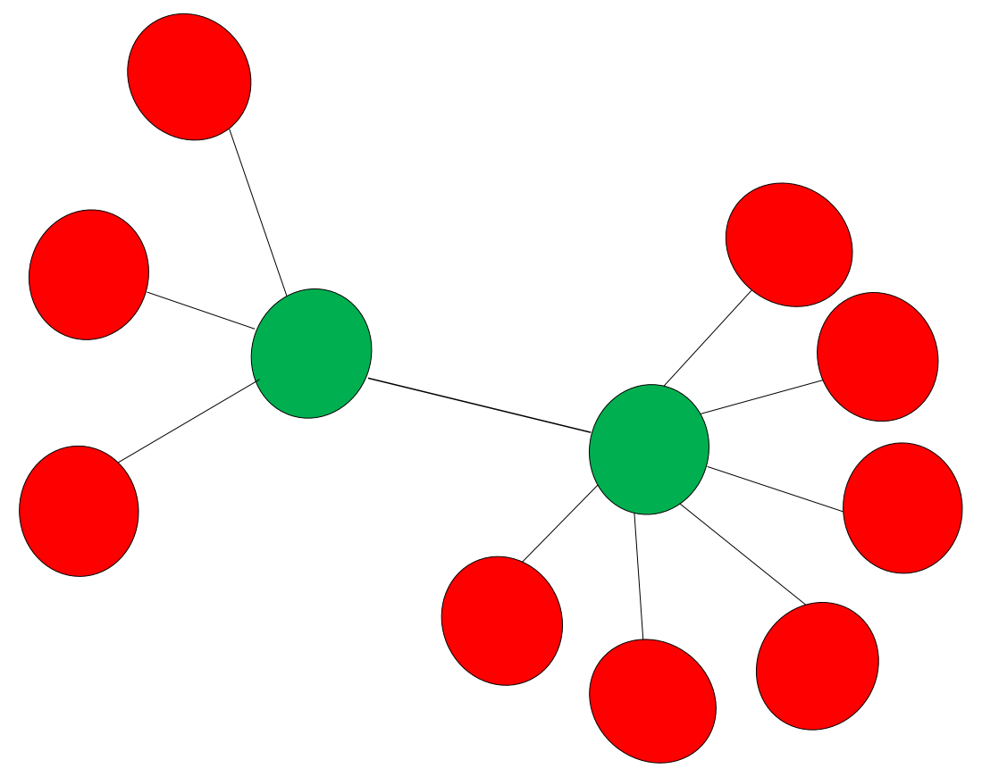


Figure 2: Extracted subgraph.

2) Extracting enclosed subgraphs.

- We extract a **subgraph** $\boldsymbol{Sub_i(t,g)^+}$ that contains the known regulatory pairs themselves and their 1-hop neighbors on a noisy skeleton GRN_i' as the positive subgraphs.
- Meanwhile, **randomly** select t ∈ TF and g ∈ {TF, G}, (t, g) ∉ L, extract a **subgraph** $\boldsymbol{Sub_i(t,g)}^-$ containing themselves and their 1-hop neighbors on the noisy skeleton GRN′i as the negative subgraphs.

3) Constructing nodes features in each subgraph.

- Two general categories of features are used to build node features:
  - Explicit features**features** are calculated using the gene expression vector E_i of gene *i*, *i* ∈{TF, G}, it is comprised of :
    - the mean (μ),
    - the standard deviation (σ),
    - the quantiles of expression values Q_0, Q_1, Q_2, Q_3 and Q4
    - minimum expression value = Q_0
    - maximum expression value = Q_4 .
  - **Structural embedding features** are represented as Graph embedding.
    - Graph embedding is a learned continuous feature representation for nodes in networks.
    - Nod2Vec.

4) Building ensemble Graph convolution networks classifiers:

- Graph convolutional network models are a graph neural network architecture type that can exploit the graph structure and convolutionally aggregate node information from the neighborhoods.
- Self-Attention Graph Pooling (SAGPool), a hierarchical graph pooling-related graph pooling method for GNN. SAGPool permits pooling with both node characteristics, and graph topology is considered.
- In SAGPool, the self-attention mechanism can differentiate between nodes that should be retained.

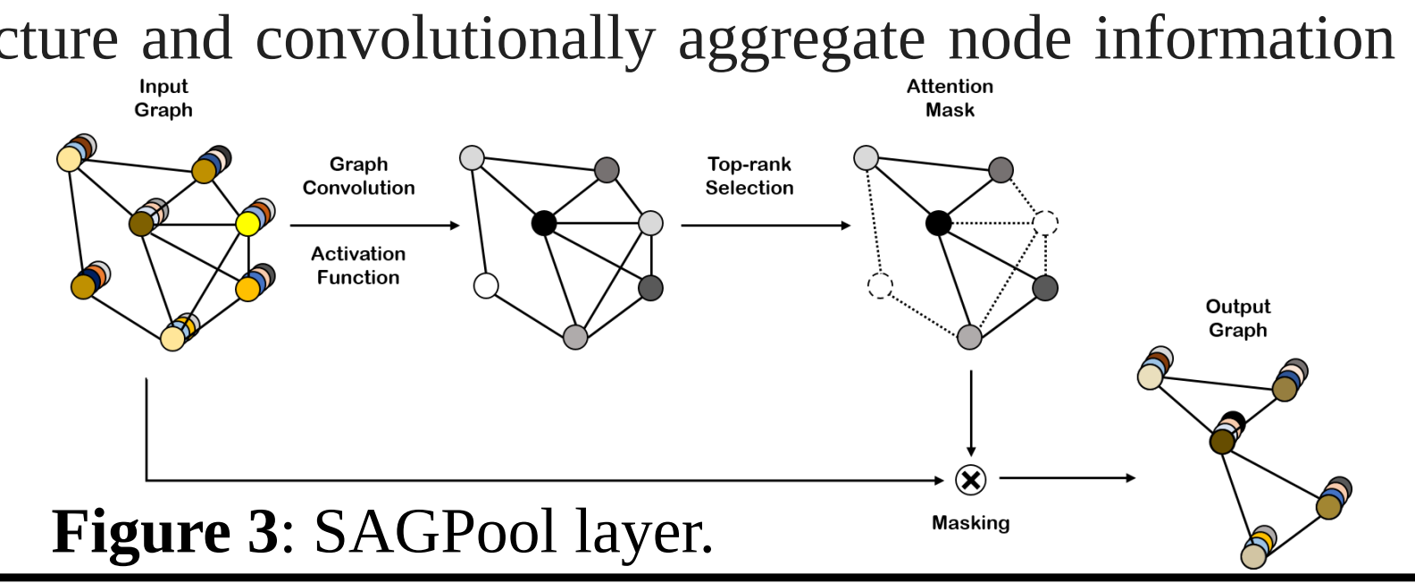
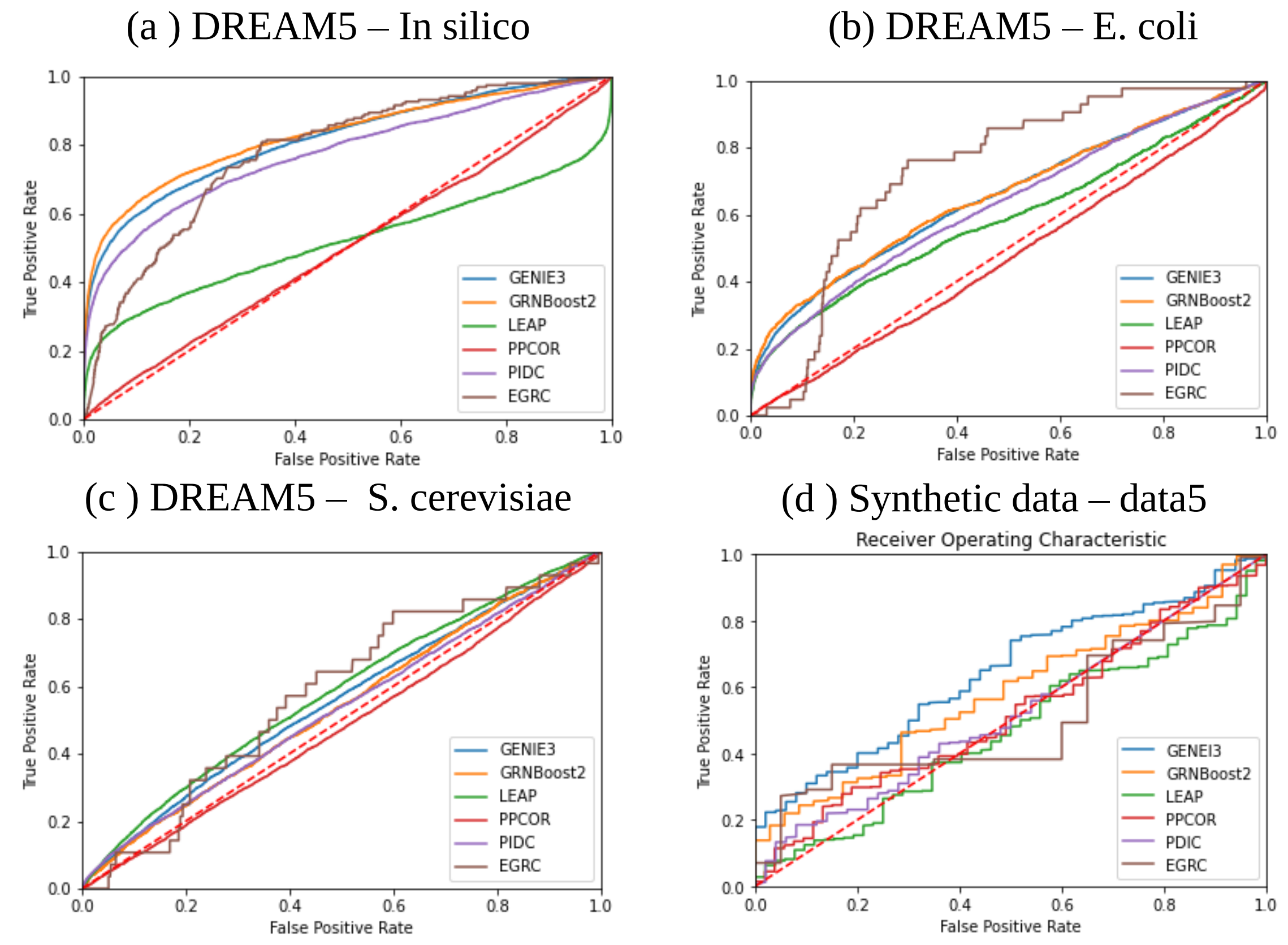


Figure 3: SAGPool layer.

Results



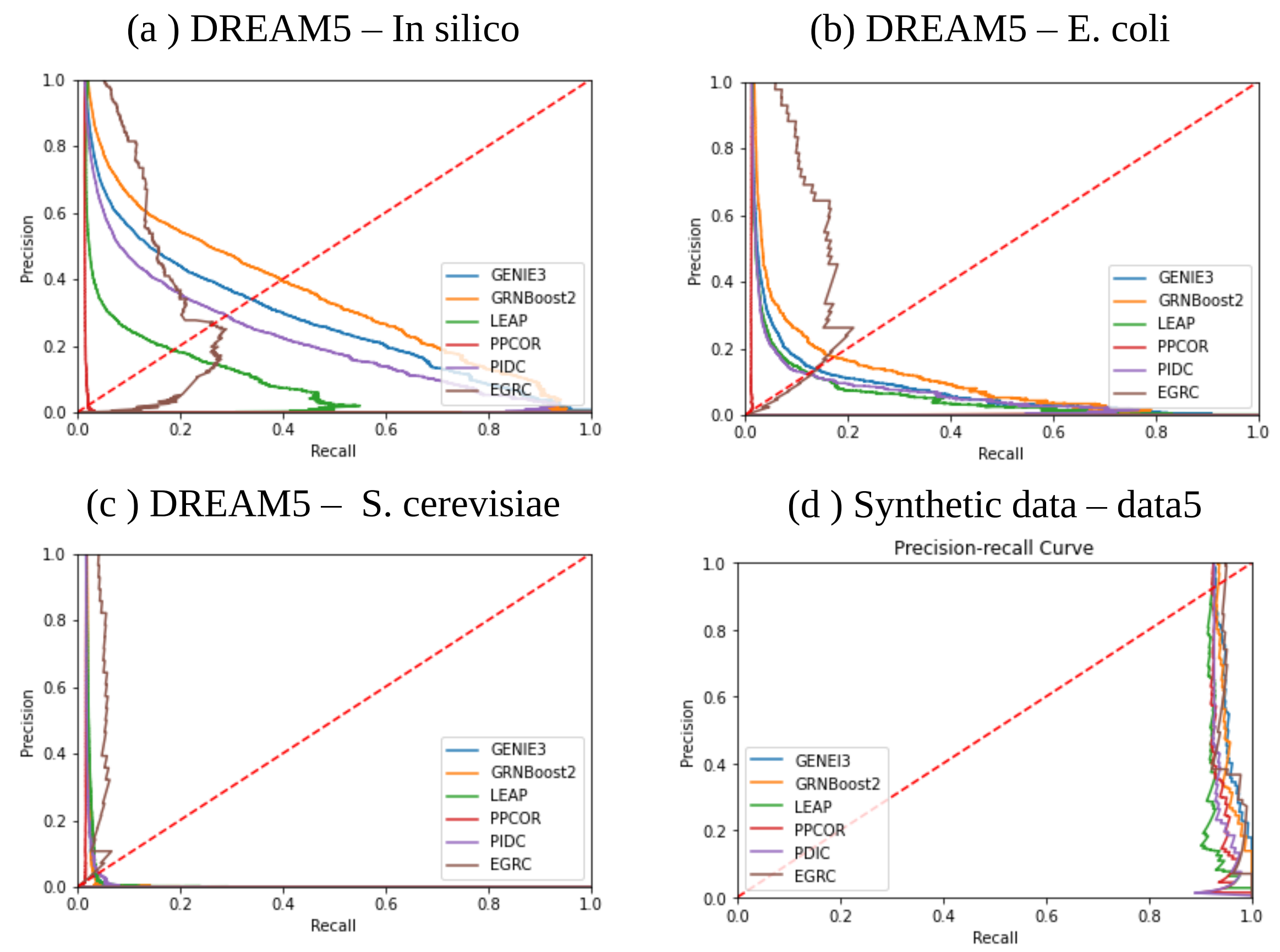
(a) DREAM5 – In silico

(b) DREAM5 – E. coli

(c) DREAM5 – S. cerevisiae

(d) Synthetic data – data5

Figure 4: Comparison of AUROC values of AGRN with other methods using DREAM5 data of (a) *in silico*, (b) *E. coli*, (c) *S. cerevisiae*, and (d) Synthetic data – data5.



(a) DREAM5 – In silico

(b) DREAM5 – E. coli

(c) DREAM5 – S. cerevisiae

(d) Synthetic data – data5

Figure 5: Comparison of AUPR values of AGRN with other methods using DREAM5 data of (a) *in silico*, (b) *E. coli*, (c) *S. cerevisiae*, and (d) Synthetic data – data5.

Table 3. Comparison of EGRC performance with the existing method using the DREAM5 datasets. The best score values are bold-faced.

Method	*In silico*		*E. coli*		*S. cerevisiae*	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
PLSNET	0.85	0.24	0.57	0.06	0.51	0.02
TIGRESS	0.75	0.29	0.58	0.06	0.51	0.02
CLR	0.77	0.25	0.59	0.08	0.52	0.02
ARACNE	0.76	0.19	0.57	0.07	0.50	0.02
MRNET	0.67	0.19	0.53	0.04	0.50	0.02
EGRC	**0.78**	**0.17**	**0.76**	**0.14**	**0.59**	**0.05**