Bioinformatics & Machine Learning Lab

AGRN: Accurate inferring Gene Regulatory Network based on ensemble machine learning methods

Duaa Mohammad Alawad, Ataur Katebi[2,3] , Md Wasi Ul Kabir, Md Tamjidul Hoque
email: dmalawad@uno.edu, a.katebi@neu.edu, mkabir4@uno.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.

THE UNIVERSITY of NEW ORLEANS

Introduction

- Biological processes are regulated by underlying genes and their interactions that form gene regulatory networks (GRNs).
- Dysregulation of these GRNs can cause complex diseases such as cancer, Alzheimer's, and diabetes.
- Accurate GRN inference is critical for elucidating gene function, allowing for the faster identification and prioritization of candidate genes for functional investigation.
- We developed a method named AGRN that infers GRNs by employing an ensemble of machine-learning algorithms.

Materials

Table 1: DREAM4 and DREAM5 Datasets.

Dataset	No. of TFs	No. of Genes
DREAM4 Network 1	100	100
DREAM4 Network 2	100	100
DREAM4 Network 3	100	100
DREAM4 Network 4	100	100
DREAM4 Network 5	100	100
DREAM5 Network (*In silico*)	195	1643
DREAM5 Network (*S. cerevisiae*)	333	5949
DREAM5 Network (*E. coli*)	334	4511

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

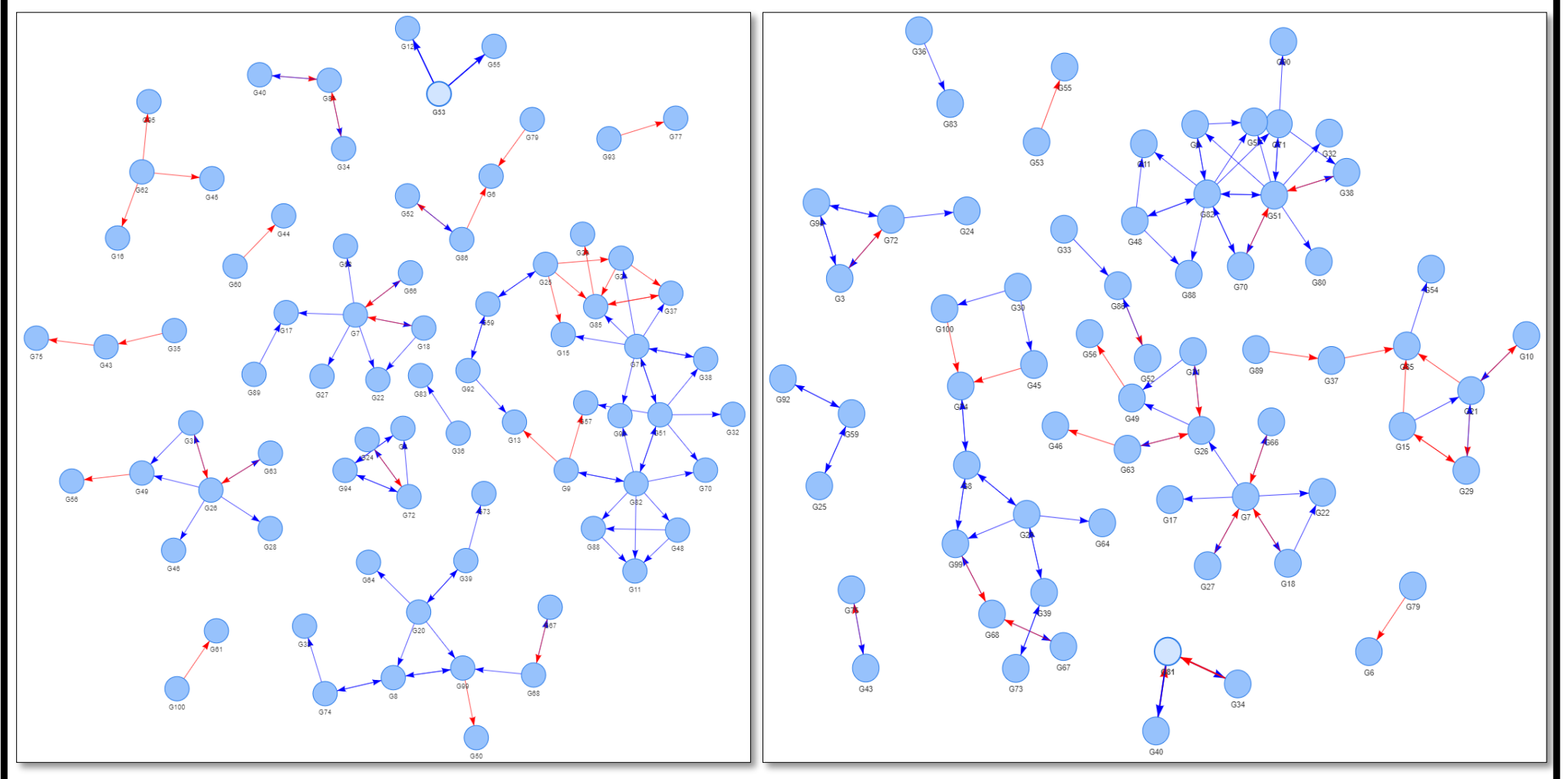
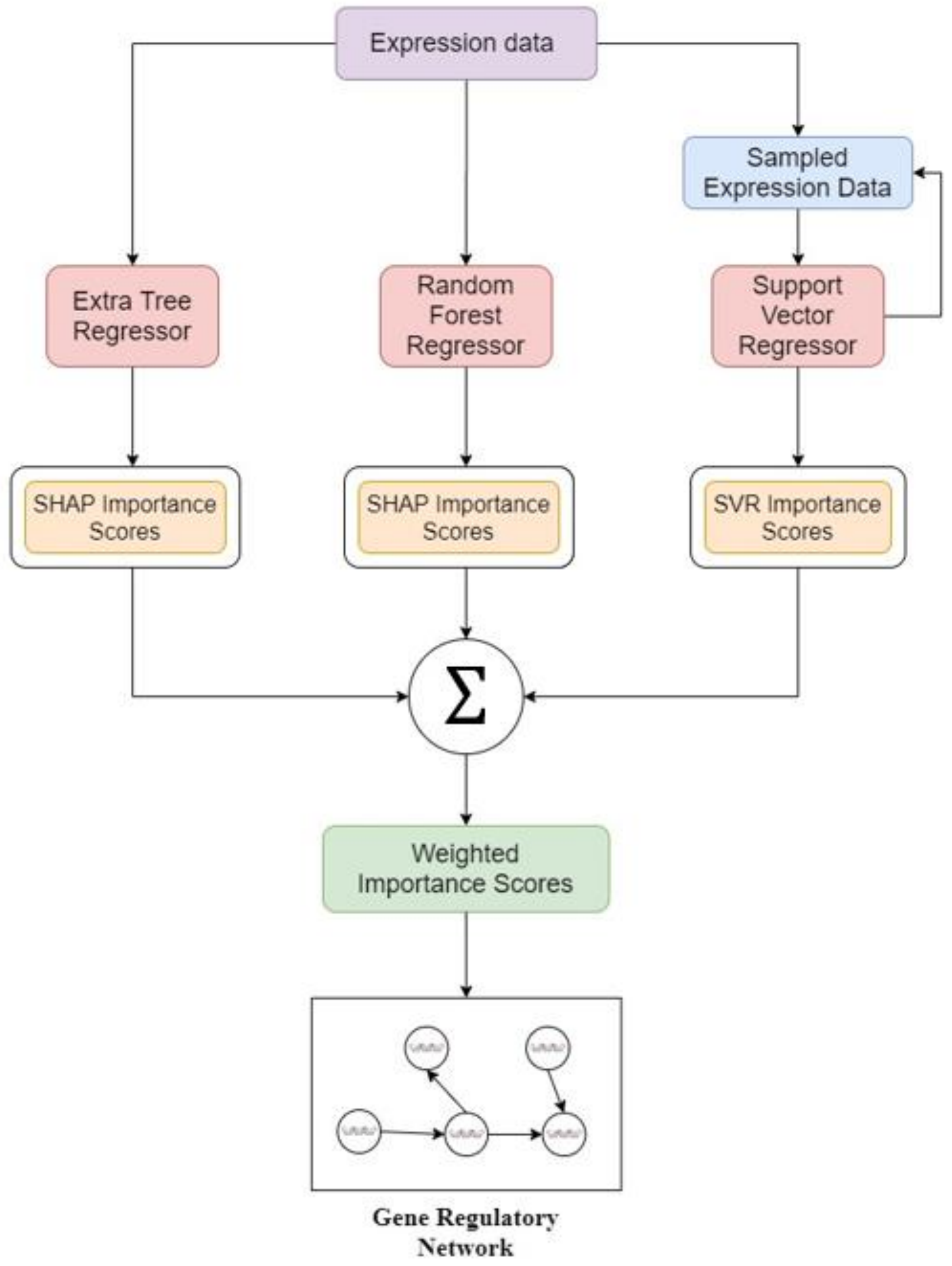


Figure 5. Predicted gene regulatory network (GRN) for network#5 of the DREAM4 dataset. The red edges represent false positives, and the blue edges represent true positives. Python library *pyvis* is used to draw the gene regulatory network. (a) GENIE3 predicted GRN. The number of blue edges is 69, whereas the number of red edges is 31. (b) AGRN predicted GRN. The number of blue edges is 72, whereas the number of red edges is 28.

Methods and results



Expression data

Sampled Expression Data

Extra Tree Regressor

Random Forest Regressor

Support Vector Regressor

SHAP Importance Scores

SHAP Importance Scores

SVR Importance Scores

Σ

Weighted Importance Scores

Gene Regulatory Network

Figure 1. The framework of the AGRN to predict the gene regulatory network.

- **RFR** and **ETR** are two tree-based regressors that consist of many decision trees to improve the prediction performance.
- We select r subsampling of observation where r is a uniformly randomly generated integer number between 20% and 80% of the observations (rows of the expression data matrix).
- The subsampling process is repeated 200 times to generate 200 models for each target gene.

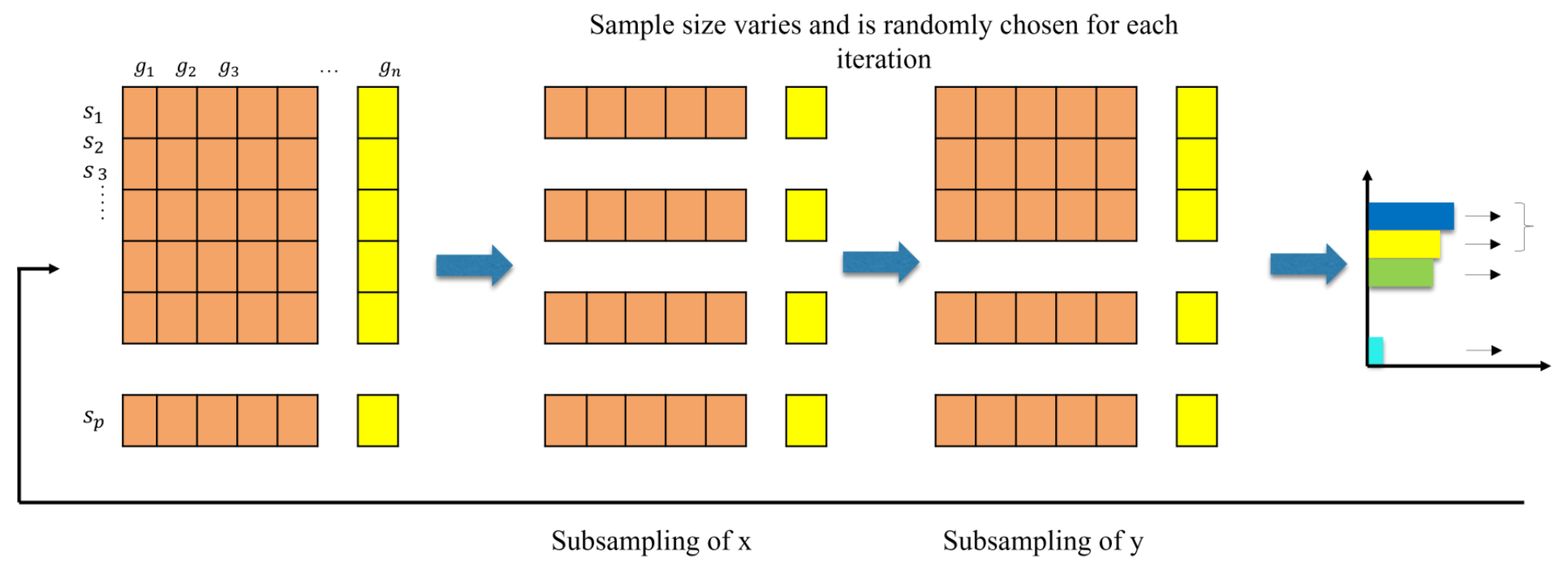


Figure 2. Ensemble SVR.

- Shapley values were used to determine how much each feature contributed to the model's final prediction

$$\phi_j\,(v) = \sum_{S\subseteq\{1,..,p)\setminus\{j\}} \frac{|S|!\ (p\ -\ |S|\ -\ 1)!}{p!}\ (v(S \cup \{j\}\,)\ -\ v\,(S))$$

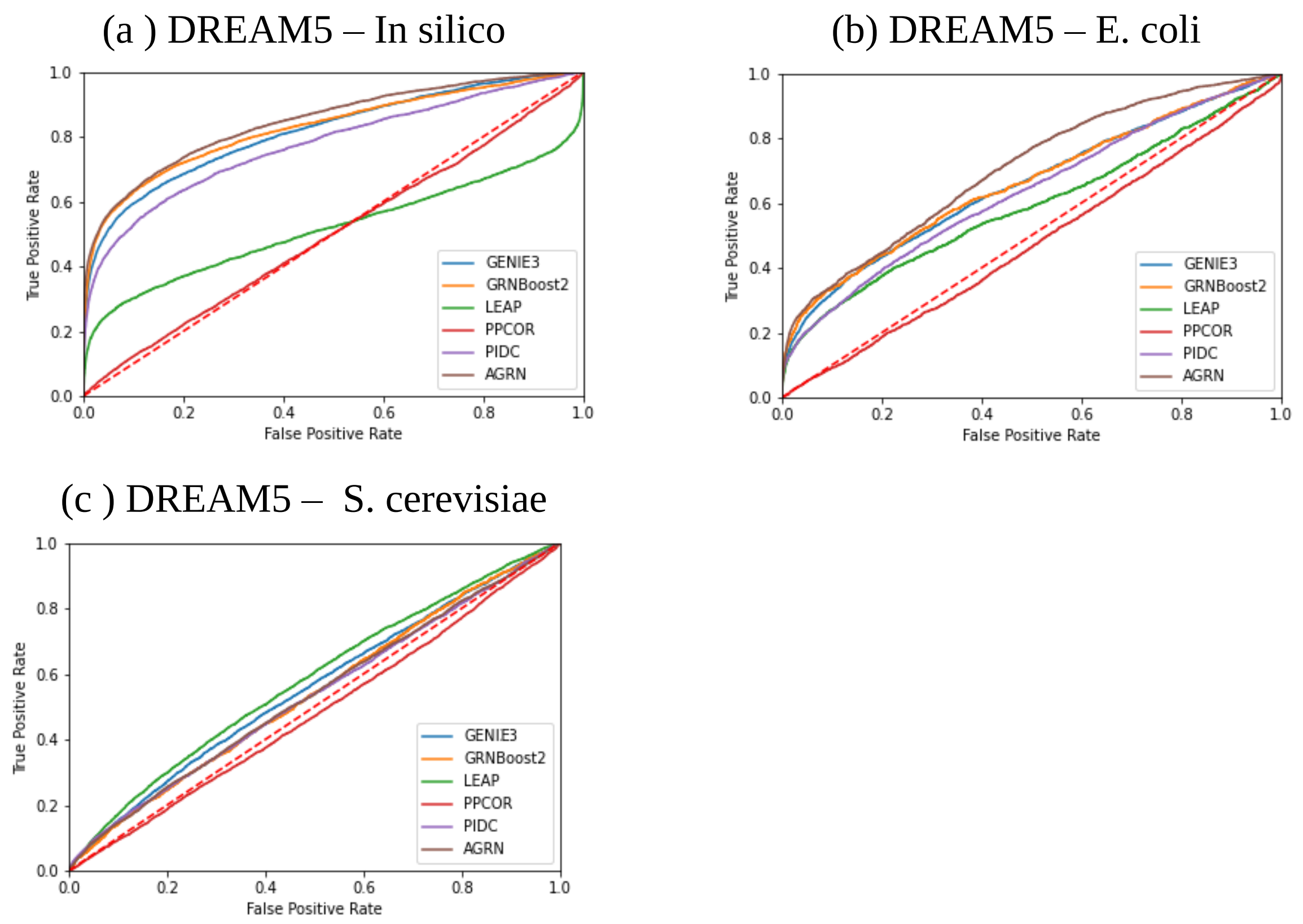
Where:
j is a feature, ϕ_j is the Shapley value for feature *j*, p is the number of features, S is the subset of features before adding the *jth* feature, *v* is the prediction value.

- After collecting the importance scores from the three methods, we take a weighted average to calculate the aggregate importance scores, as shown in Equation.

$$\mathcal{R}_g = \frac{\omega_1 *\ (\phi_g(R))\ + \omega_2 *\ (\phi_g(E)) +\ \ \omega_3 *\ (C_g(SV))}{\omega_1 + \omega_2 + \omega_3}$$

.

- We applied a **grid search** technique to find the optimal weights to calculate the final importance scores.

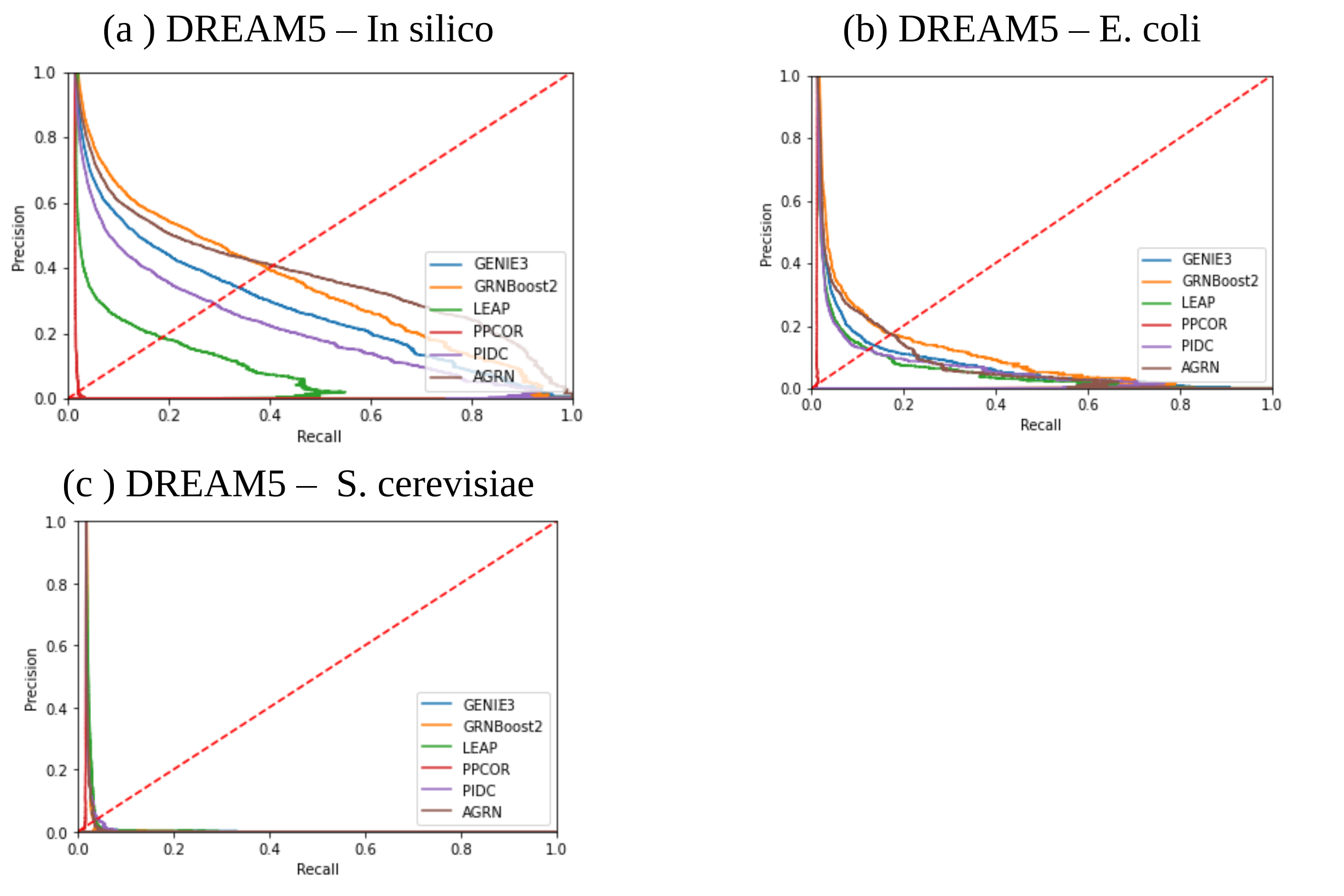


(a) DREAM5 – In silico

(b) DREAM5 – E. coli

(c) DREAM5 – S. cerevisiae

Figure 3: Comparison of AUROC values of AGRN with other methods using DREAM5 data of (a) in silico, (b) E. coli, and (c) S. cerevisiae.



(a) DREAM5 – In silico

(b) DREAM5 – E. coli

(c) DREAM5 – S. cerevisiae

Figure 4: Comparison of AUPR values of AGRN with other methods using DREAM5 data of (a) in silico, (b) E. coli, and (c) S. cerevisiae.

Conclusion

- By combining three machine learning algorithms, we developed an ensemble machine learning method named AGRN that infers GRNs using the importance scores.
- In AGRN, we combined the importance scores calculated in each of RFR and ETR based on their Shapley values.
- In addition, importance scores were calculated from multiple SVR models with iterative sampling. Moreover, we optimize the SVR hyperparameters and use the weighted average of the three methods (Shap-BasedOnRFR, ShapBasedOnETR, SVR) to have the final importance scores.
- The comparison of AGRN with five benchmark-ing methods (GRNBoost2, PPCOR, GENIE3, LEAP, and PIDC) using five networks from DREAM4 dataset and two datasets from DREAM5 shows that AGRN outperforms the other methods.
- For example, using the *In silico* data from DREAM5, compared with the second-ranked method (GRNBoost2), AGRN achieves an improvement of 2.42% and 8.83% based on AUROC and AUPR scores, respectively. Al-so, using *E. coli* dataset, the comparison shows that AGRN achieves an improvement of 7.65% and 7.14% based on the AU-ROC and AUPR scores, respectively.
- Therefore, these results allow us to conclude that, rather than using a single importance score, AGRN can improve performance on GRN inference by combining importance scores from SHAP-based RFR, SHAP-based ETR, and optimized SVR.