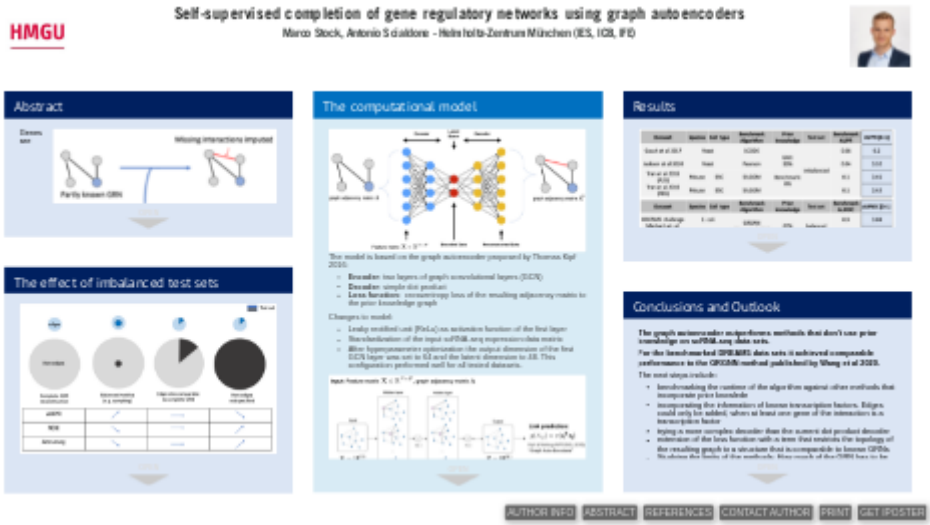
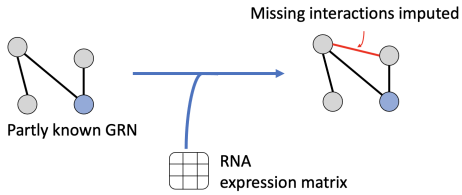
Self-supervised completion of gene regulatory networks using graph autoencoders



Marco Stock, Antonio Scialdone - Helmholtz-Zentrum München (IES, ICB, IFE)

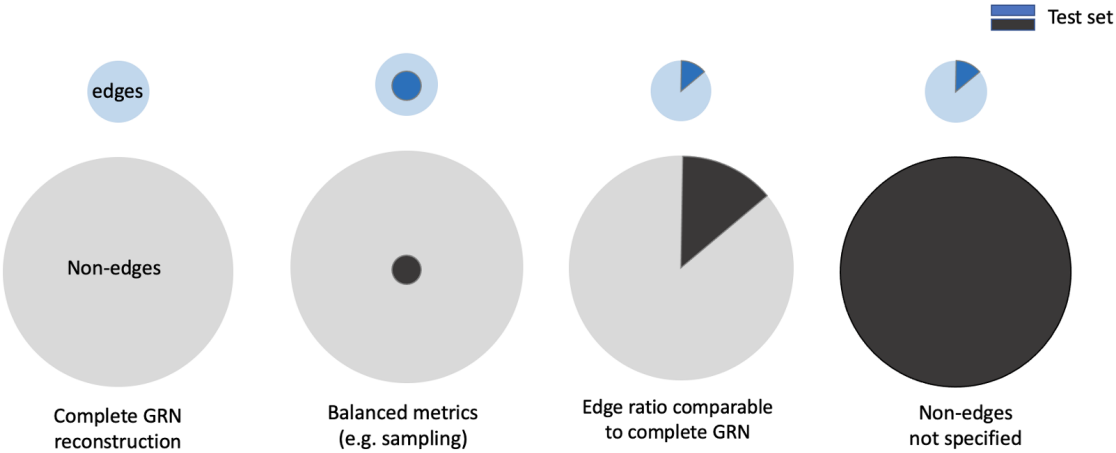
PRESENTED AT:

ABSTRACT



Genes are interacting in complex network structures. The computational reconstruction of such Gene Regulatory Networks (GRN) in the representation of graphs from sequencing data, remains a very important and challenging task. While individual gene interactions may be experimentally detected and validated by, e.g., knockout experiments, this approach lacks the scalability to infer large interaction networks. The recent research progress on Graph Neural Networks (GNN) enabled their successful application in several problems, such as in the protein folding predictor AlphaFold 2. Here we use a Variational Graph Autoencoder (VGAE) to complete partly known GRNs borrowing information from RNA sequencing data sets. This self-supervised machine learning method uses a given incomplete GRN to predict missing gene interactions. The predicted gene interactions can then be validated experimentally. The approach is suitable for both bulk and single cell RNA-sequencing data combined with partly known ground truth interaction networks. In our ongoing work, the first version of the model is applied to different data sets to get an unbiased performance estimate of the predictions and it is moreover benchmarked against other supervised methods of gene interaction inference.

THE EFFECT OF IMBALANCED TEST SETS



AUPR	↗	→	↘
ROC	↘	→	↗
Accuracy	↘	→	↗

Different evaluation metrics are used in papers.

Not every metric is suited for highly imbalanced test sets as GRNs.

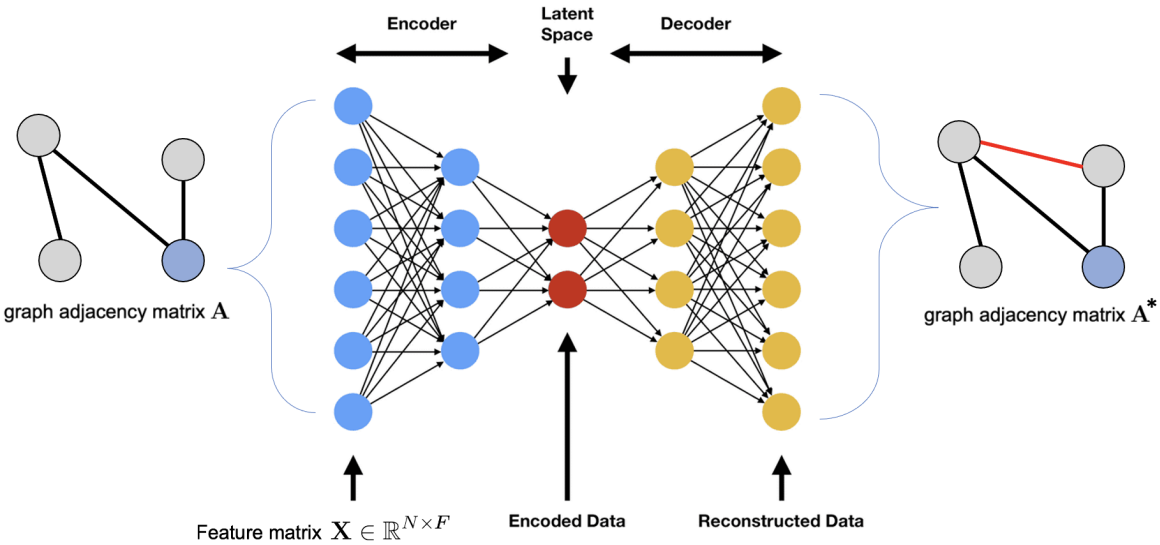
Three positive/negative ratios of the test sets can be distinguished:

- **original GRN (ratio ~1:300):**
  used in inference of complete GRNs without prior knowledge.
  Higher ROC and accuracy expected.

- **balanced test set (ratio 1:1):**
  artifically sampled test sets.
  Higher AUPR expected.

- **inference of new edges (ratio ~1:30.000):**
  realistic scenario when inferring additional edges and prior knowledge consists of positive edges only.
  Low AUPR expected.

THE COMPUTATIONAL MODEL

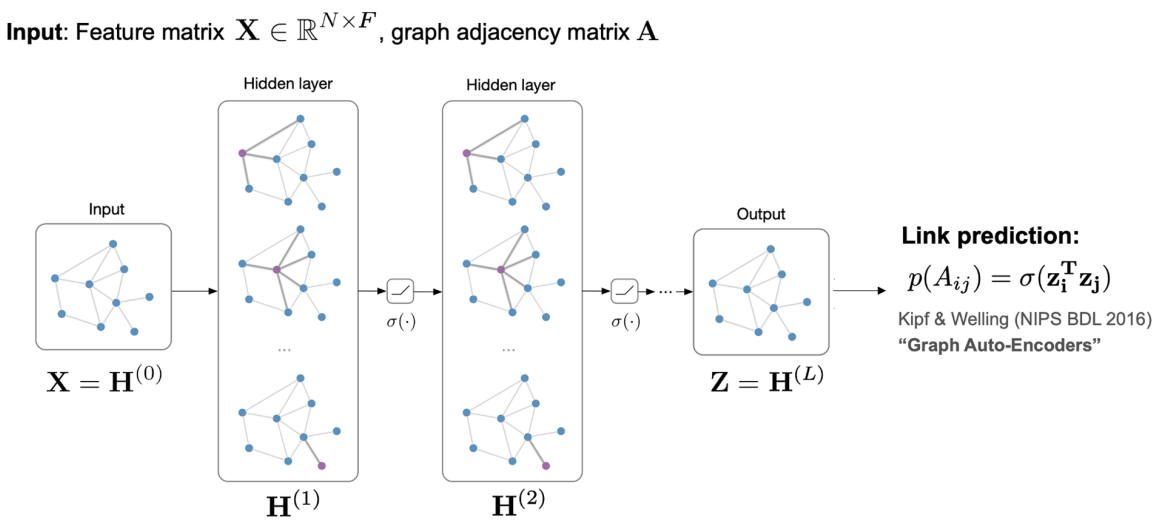


The model is based on the graph autoencoder proposed by Thomas Kipf 2016:

- **Encoder**: two layers of graph convolutional layers (GCN)
- **Decoder**: simple dot product
- **Loss function**: crossentropy loss of the resulting adjacency matrix to the prior knowledge graph

Changes to model:

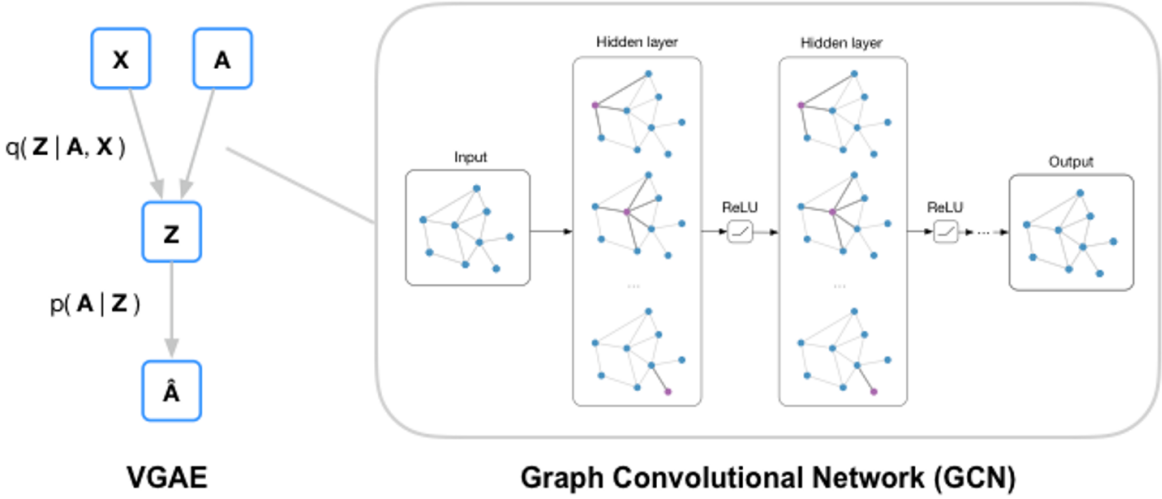
- Leaky rectified unit (ReLu) as activation function of the first layer
- Standardization of the input scRNA-seq expression data matrix
- After hyperparameter optimization the output dimension of the first GCN layer was set to 64 and the latent dimension to 48. This configuration performed well for all tested datasets.



<u>Variational autoencoder</u>

- adds additional regularization by learning a probablitiy distribution in latent space and drawing samples from the distribution for reconstruction
- loss function is extended by a Kullback-Leibler (KL) divergence on the distribution of the latent space variables Z.

For stabilizing training a warmup function for the KL loss was added, as suggested by Sønderby et al 2016 and Bowman et al 2016.



VGAE **Graph Convolutional Network (GCN)**

RESULTS

Dataset	Species	Cell type	Benchmark Algorithm	Prior knowledge	Test set	Benchmark AUPR	AUPR [0-1]
Gasch et al 2017	Yeast		SCODE	GAE: 99% Benchmark: 0%	imbalanced	0.06	0.2
Jackson et al 2020	Yeast		Pearson			0.04	0.32
Tran et al 2019 (A2S)	Mouse	ESC	SILGGM			0.1	0.45
Tran et al 2019 (FBS)	Mouse	ESC	SILGGM			0.1	0.42

Dataset	Species	Cell type	Benchmark Algorithm	Prior knowledge	Test set	Benchmark AUROC	AUROC [0-1]
DREAM5 challenge Marbach et. al. 2012	E. coli		GRGNN Wang et al 2020	67%	balanced	0.9	0.88
	Yeast					0.88	0.85

Tested on six scRNA-Seq datasets, wich were provided with associated gold standard ground thruth networks by two papers:

1. **Stone et al 2021 Benchmarking paper for scRNA-Seq GRN inference:**
   The positive/negative ratio of the test set was chosen to match the original GRN ratio.
2. **DREAM5 network inference challenge:**
   The test sets were balanced to be able to compare performance to the numbers of the GRGNN algorithm by Wang et al 2020.
   The numbers reported represent 3-fold crossvalidation.

The training of the variational auto encoder was not able to converge for all training sets. Therefore the numbers reported are the scores of the regular graph autoencoder structure without the KL loss and latent space sampling.

An overview of the datasets is provided in the following figure.

Dataset	Technology	Species	Cell type	# cells	# genes
Gasch et al 2017	scRNA-Seq	Yeast		163	3.847
Jackson et al 2020		Yeast		17.396	5.736
Tran et al 2019 (A2S)		Mouse	ESC	2.369	6.618
Tran et al 2019 (FBS)		Mouse	ESC	3.324	6.621
DREAM5 challenge Marbach et. al. 2012	microarrays	E. coli		805	4.511
		Yeast		536	5.950

CONCLUSIONS AND OUTLOOK

The graph autoencoder outperforms methods that don't use prior knowledge on scRNA-seq data sets.

For the benchmarked DREAM5 data sets it achieved comparable performance to the GRGNN method published by Wang et al 2020.

The next steps include:

- benchmarking the runtime of the algorithm against other methods that incorporate prior knowlede
- incorporating the information of known transcription factors. Edges could only be added, when at least one gene of the interaction is a transcription factor
- trying a more complex decoder than the current dot product decoder
- extension of the loss function with a term that restricts the topology of the resulting graph to a structure that is comparable to known GRNs
- Studying the limits of the methods: How much of the GRN has to be known in advance to get reasonable reconstruction results? How many variations in the cells are needed produce reasonable evalution scores?

Acknowledgement:

Marco Stock is supported by the Helmholtz Association under the joint research school "Munich School for Data Science - MUDS".

References:

T. N. Kipf, M. Welling, Variational Graph Auto-Encoders, NIPS Workshop on Bayesian Deep Learning (2016)

C. K. Sonderby et al, Ladder Variational Autoencoders (2016)

S. R. Bowman et al, Generating Sequences from a Continous Space, SIGNLL Conference on Computational Natural Language Learning (2016)

M. Stone et al, Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data (2021)

J. Wang et al, Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks, Computational and structural biotechnology journal (2020)

AUTHOR INFO

Marco Stock (1,2,3), Antonio Scialdone (1,2,3)

(1) Helmholtz Zentrum München Institute of Epigenetics and Stem cells (IES)

(2) Helmholtz Zentrum München Institute of Computational Biology (ICB)

(3) Helmholtz Zentrum München Institute of Functional Epigenetics

ABSTRACT

Genes are interacting in complex network structures. The computational reconstruction of such Gene Regulatory Networks (GRN) in the representation of graphs from sequencing data, remains a very important and challenging task. While individual gene interactions may be experimentally detected and validated by, e.g., knockout experiments, this approach lacks the scalability to infer large interaction networks. The recent research progress on Graph Neural Networks (GNN) enabled their successful application in several problems, such as in the protein folding predictor AlphaFold 2. Here we use a Variational Graph Autoencoder (VGAE) to complete partly known GRNs borrowing information from RNA sequencing data sets. This self-supervised machine learning method uses a given incomplete GRN to predict missing gene interactions. The predicted gene interactions can then be validated experimentally. The approach is suitable for both bulk and single cell RNA-sequencing data combined with partly known ground truth interaction networks. In our ongoing work, the first version of the model is applied to different data sets to get an unbiased performance estimate of the predictions and it is moreover benchmarked against other supervised methods of gene interaction inference.

REFERENCES

T. N. Kipf, M. Welling, Variational Graph Auto-Encoders, NIPS Workshop on Bayesian Deep Learning (2016)

C. K. Sonderby et al, Ladder Variational Autoencoders (2016)

S. R. Bowman et al, Generating Sequences from a Continous Space, SIGNLL Conference on Computational Natural Language Learning (2016)

M. Stone et al, Identifying strengths and weaknesses of methods for computational network inference from single cell RNA-seq data (2021)

J. Wang et al, Inductive inference of gene regulatory network using supervised and semi-supervised graph neural networks, Computational and structural biotechnology journal (2020)