



# **miDGD: Prediction of miRNA activity levels based on gene expression data using Deep Generative Decoder model**

Master's in Bioinformatics Thesis

**Farhad Zamani**

Supervisor:

Jakob Skou Pedersen

Spring 2024

**miDGD: Prediction of miRNA activity levels  
based on gene expression data using Deep  
Generative Decoder model**

Master's in Bioinformatics Thesis

Farhad Zamani

Supervisor: Jakob Skou Pedersen

Spring 2024





Aarhus University

<http://www.au.dk>

# AARHUS UNIVERSITY

**Thesis title:**

miDGD: Prediction of miRNA activity levels based on gene expression data using Deep Generative Decoder model

**Thesis Period:**

Spring Semester 2024

**Author:**

Farhad Zamani

**Supervisor:**

Jakob Skou Pedersen

**Page Numbers:** 64

**Date of Completion:**

June 15, 2024

**Master's of Science in Bioinformatics**

Bioinformatics Research Center (BiRC)

10 Aarhus University

Denmark

**In collaboration with:**

Department of Molecular Medicine (MOMA)

15 Aarhus University Hospital

Denmark



# Abstract

## Motivation

20 microRNAs play an important role in regulating gene expression at a posttranscriptional level. In cancer cells, miRNAs are often dysregulated, acting as either oncomiRs or tumor suppressors. Therefore, understanding miRNA regulation is important in cancer research and potentially elsewhere. However, there are limitations to studying miRNAs in single-cell RNA sequencing (scRNA-seq) settings.

25 The state-of-the-art to infer miRNA expression levels involves several computational approaches that leverage the relationship between miRNAs and their target mRNAs, including motif enrichment analysis and machine learning models like XGBoost. This thesis aims to explore and evaluate the ability of generative AI approaches to improve the prediction of miRNA expression levels from gene expression data in bulk RNA-seq and scRNA-seq settings.

30

## Results

We present the miDGD, a Deep Generative Decoder (DGD) model that can infer miRNA activity levels based on only gene expression data. The miDGD model learns the shared representation of gene and miRNA expression and handles complex parameterized latent distributions. The result shows that miDGD model can be used to predict miRNA expression in bulk RNA-seq and sparse data equivalent to scRNA-seq experiments.



## 40 Acknowledgements

First and foremost, I would like to express my sincere gratitude to everyone who has supported me throughout the completion of this thesis.

I am deeply grateful to my supervisor, Professor Jakob Skou Pedersen, for your invaluable guidance, continuous support, and patience during my research. Your  
45 immense knowledge and experience have been influential in shaping this thesis. Your enthusiasm for the research is truly inspiring.

I would also like to thank the members of the Skou Pedersen Group for the insightful discussions and help through the thesis, especially Mathilde Diekema for your valuable input and for reviewing my thesis, and Asta Mannstaedt Rasmussen  
50 for the assistance in navigating and working with the dataset. Your contributions have significantly improved the quality of my work.

Special thanks to the Bioinformatics Research Center (BiRC) and the Molecular Medicine Department (MOMA) at Aarhus University for providing a stimulating environment. I am particularly thankful to my fellow Master's in Bioinformatics  
55 students for their moral support and the fun times we shared.

I would like to extend my heartfelt thanks to my family for their unwavering support and encouragement. To my parents, thank you for your endless love and for believing in me. I am also grateful to my friends in PPI Denmark for their support and for making this journey enjoyable.





## 60 **Abbreviations**

**ACC** Adrenocortical Carcinoma.

**BLCA** Bladder Urothelial Carcinoma.

**BRCA** Breast Invasive Carcinoma.

**CESC** Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma.

65 **CHOL** Cholangiocarcinoma.

**COAD** Colon Adenocarcinoma.

**DGD** Deep Generative Decoder.

**DLBC** Lymphoid Neoplasm Diffuse Large B-cell Lymphoma.

**ESCA** Esophageal Carcinoma.

70 **GMM** Gaussian Mixture Model.

**HNSC** Head and Neck Squamous Cell Carcinoma.

**KICH** Kidney Chromophobe.

**KIRC** Kidney Renal Clear Cell Carcinoma.

**KIRP** Kidney Renal Papillary Cell Carcinoma.

75 **LAML** Acute Myeloid Leukemia.

**LGG** Brain Lower Grade Glioma.

**LIHC** Liver Hepatocellular Carcinoma.

**LUAD** Lung Adenocarcinoma.

**LUSC** Lung Squamous Cell Carcinoma.

80 **MESO** Mesothelioma.

**OV** Ovarian Serous Cystadenocarcinoma.

**PAAD** Pancreatic Adenocarcinoma.

**PCPG** Pheochromocytoma and Paraganglioma.

**PRAD** Prostate Adenocarcinoma.

**READ** Rectum Adenocarcinoma. 85

**SARC** Sarcoma.

**scRNA-seq** Single-cell RNA sequencing.

**SKCM** Skin Cutaneous Melanoma.

**STAD** Stomach Adenocarcinoma.

**TCGA** The Cancer Genome Atlas. 90

**TGCT** Testicular Germ Cell Tumors.

**THCA** Thyroid Carcinoma.

**THYM** Thymoma.

**UCEC** Uterine Corpus Endometrial Carcinoma.

**UCS** Uterine Carcinosarcoma. 95

**UTR** Untranslated Region.

**UVM** Uveal Melanoma.

# Contents

100	<b>Abbreviations</b>	ix
	<b>1 Introduction</b>	1
	1.1 Background	1
	1.2 Thesis objectives	2
	1.3 Scope and delimitations	3
105	1.4 Thesis Outline	3
	<b>2 Related Works</b>	5
	2.1 miRNA inference methods	5
	2.2 Deep Generative Decoder model	6
	<b>3 Background Theory</b>	9
110	3.1 miRNA Biology	9
	3.1.1 miRNA Biogenesis	9
	3.1.2 miRNA as post-transcriptional regulator	11
	3.1.3 miRNA on Cancer Research	11
	3.2 RNA Sequencing	12
115	3.3 Deep Learning	13
	3.3.1 Variational Autoencoders	13
	3.3.2 Deep Generative Decoder	14
	3.3.3 Deep Learning	16
	<b>4 Methods</b>	19
120	4.1 Dataset	19
	4.1.1 Preprocessing	19
	4.1.2 Train, validation, test split	20
	4.2 Model Overview	20
	4.2.1 Model and training objectives	21
125	4.2.2 Representation	22
	4.2.3 Gaussian Mixture Model	22
	4.2.4 Decoder	24
	4.2.5 Prediction of new samples	24

4.3 Experiments . . . . .	25	
4.3.1 Materials . . . . .	26	130
4.3.2 Cross-validation . . . . .	26	
4.3.3 Parameter initialization and hyperparameter tuning . . . . .	27	
4.3.4 DGD with one modality as sanity checks . . . . .	27	
4.3.5 miDGD for miRNA expression levels prediction based on gene expression data . . . . .	28	135
4.3.6 miRNA prediction using downsampled gene expression data . . . . .	28	
4.3.7 Performance evaluation . . . . .	28	
4.4 Code Availability . . . . .	29	
<b>5 Results and Discussion</b>	<b>31</b>	
5.1 Brief overview of the data . . . . .	31	140
5.1.1 Samples metadata overview . . . . .	31	
5.1.2 Exploring Dimensionality Reduction in Dataset . . . . .	33	
5.2 DGD model trained on only mRNA or miRNA independently as sanity checks . . . . .	34	
5.2.1 Model training . . . . .	34	145
5.2.2 Model performance . . . . .	35	
5.3 miDGD to predict miRNA expression level based on gene expres- sion data . . . . .	36	
5.3.1 Hyperparameter tuning . . . . .	37	
5.3.2 Training metrics . . . . .	38	150
5.3.3 Model performance . . . . .	39	
5.4 miDGD performance to predict tissue-specific miRNA activity . . . . .	43	
5.5 Clustering the representation to the corresponding GMM compo- nent and primary site . . . . .	46	
5.6 Evaluating miDGD performance in predicting miRNA expression level in downsampled datasets . . . . .	48	155
5.7 Benchmarking with miRSCAPE . . . . .	54	
5.8 Discussion and Future Perspective . . . . .	54	
<b>6 Conclusion</b>	<b>57</b>	
<b>References</b>	<b>59</b>	160
<b>A Appendix A: Formula</b>	<b>63</b>	
A.1 Gaussian Mixture Model . . . . .	63	