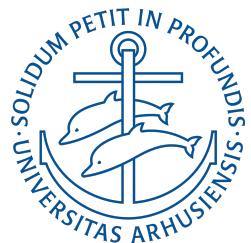


# **Factorizing representations of cancer mutations with a Deep Generative Decoder**

**Cancer type prediction using generative AI**



**Astrid Klitgaard Dahl**

**Supervisor:** Jakob Skou Pedersen

Bioinformatics Research Center (BiRC)

Aarhus University

In collaboration with:

Department of Molecular Medicine (MOMA)

Aarhus University Hospital

This thesis is submitted for the degree of

*Master of Science in Bioinformatics*

June 2025



## Acknowledgements

I would like to thank my main supervisor Professor Jakob Skou Pedersen for many insightful discussions and guidance on the subjects of this thesis. I also thank my daily supervisor Mathilde Diekema for helping with the scope of the project and reviewing my thesis, and Farhad Zamani for helping me navigate the DGD implementation. Finally, I am grateful for the fun environment created by my fellow Master's students in the research group.



## Abstract

Understanding the complex mutational landscape of cancer requires models that can capture both the biological variability of tumor samples and the contextual factors influencing mutation patterns. In this thesis, we introduce the Factor Deep Generative Decoder (Factor DGD), a novel model that learns a latent representation factorized into two parts: one representing tumor samples and another representing mutation contexts. This approach extends traditional methods by enabling nonlinear modeling of mutation data from primary tumor biopsies in the PCAWG dataset.

We implement the Factor DGD using 5-mer mutation contexts and evaluate its ability to cluster tumor types and capture mutation signature structure. While the model demonstrates promising clustering performance, challenges remain in optimizing the Gaussian Mixture Models used to represent latent components. Compared to established approaches like mutational signatures, the Factor DGD offers a flexible framework with potential for improved biological interpretability.



# Table of contents

<b>Abbreviations</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Thesis aims . . . . .	2
1.3 Delimitations . . . . .	2
1.4 Related work . . . . .	2
<b>2 Theory</b>	<b>5</b>
2.1 Cancer development . . . . .	5
2.2 Mutational signatures . . . . .	5
2.3 Deep learning . . . . .	8
2.4 Generative modelling . . . . .	9
2.4.1 Deep Generative Decoder . . . . .	10
<b>3 Methods and data</b>	<b>13</b>
3.1 Dataset . . . . .	13
3.2 Model architecture . . . . .	14
3.2.1 Model complexity . . . . .	15
3.3 Training . . . . .	16
3.3.1 Loss functions . . . . .	16
3.3.2 Standard DGD . . . . .	17
3.3.3 Factor DGD . . . . .	17
3.3.4 Optimizer . . . . .	18
3.4 Testing . . . . .	19
3.4.1 Standard DGD . . . . .	19
3.4.2 Factor DGD . . . . .	19
3.5 Performance evaluation . . . . .	19
3.6 Code availability . . . . .	21

<b>4 Results</b>	<b>23</b>
4.1 Standard DGD . . . . .	23
4.2 Factor DGD without one-hot encoding . . . . .	24
4.3 Factor DGD with one-hot encoding . . . . .	26
4.4 Performance comparison . . . . .	30
<b>5 Discussion</b>	<b>39</b>
<b>6 Conclusion</b>	<b>43</b>
<b>References</b>	<b>45</b>
<b>Appendix A</b>	<b>47</b>
A.1 . . . . .	47

# Abbreviations

**AE** Autoencoder.

**ARI** Adjusted Rand Index.

**cfDNA** Cell-Free DNA.

**ctDNA** Circulating Tumor DNA.

**DBS** Double-Base Substitution.

**DGD** Deep Generative Decoder.

**DNA** Deoxyribonucleic Acid.

**DNN** Deep Neural Network.

**GMM** Gaussian Mixture Model.

**indels** insertions and deletions.

**KNN** K-Nearest Neighbor.

**MAP** Maximum A Posteriori.

**Mbp** Mega base pair.

**MNV** Multiple Nucleotide Variant.

**NB** Negative Binomial.

**NMF** Non-negative Matrix Factorization.

**PC** Principal Component.

**PCA** Principal Component Analysis.

**PCAWG** Pan-Cancer Analysis of Whole Genomes.

**RNA** Ribonucleic Acid.

**SBS** Single Base Substitution.

**scDGD** Single-Cell DGD.

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

**SNV** Single Nucleotide Variant.

**SV** Structural Variant.

**UV** Ultra Violet.

**VAE** Variational Autoencoder.

**WGS** Whole Genome Sequencing.