

Master's in Bioinformatics Thesis

Deep Self-Supervised Modeling of cfDNA Fragmentomics for Tumor Detection

Author:

Ruiqi Zhang

Supervisor:

Jakob Skou Pedersen



Computational Genomics & Transcriptomics Group

Department of Molecular Medicine

AARHUS UNIVERSITY

Jul 2026

Table of Contents

Acknowledgements.....	5
Abbreviations	6
1 Introduction.....	8
1.1 Thesis Objectives.....	9
1.2 Scope and Limitations	9
1.3 Thesis Outline.....	10
2 Background Theory and Related Works	11
2.1 DNA Organization and Gene Regulation.....	11
2.2 Nucleosome Positioning and Chromatin Accessibility.....	12
2.3 Liquid Biopsy	14
2.4 Cell-Free DNA.....	14
2.5 CfDNA Fragmentomics	15
2.6 Transcription Start Site	18
2.7 Machine Learning and Deep Learning.....	19
2.7.1 Self-supervised Learning	19
2.7.2 Logistic Regression and Ridge Regression.....	19
2.7.3 Deep Learning	20
Deep Learning Fundamentals.....	20
2.7.4 Transformer	21
2.7.5 Autoencoding	22
2.7.6 Masked Autoencoders.....	22
3 Methods and Data	24
3.1 Datasets	24
3.1.1 CfDNA Data.....	24
Fragment filtering and GC correction	27
3.1.2 RNA-sequencing data	28
3.2 Overall Study Design	29
3.3 Self-supervised model	31
3.3.1 Train and Validation Split.....	31
3.3.2 Model Architecture	31
3.3.3 Loss Function	32
3.4 Supervised Prediction Model	34
3.4.1 Embedding	34
3.5 Evaluation Metrics	36
3.6 Materials	36
3.7 Hyperparameters	37
3.8 Code Availability.....	37
4 Results.....	38
4.1 Exploratory Analysis	38
4.2 Self-supervised Model	40
4.2.1 Evaluation of Model Training.....	40
4.2.2 Model Interpretability Analysis.....	44

4.3 Supervised Model for Prediction 45

 4.3.1 Evaluating Tumor Status Classification Performance 45

 4.3.2 Evaluating Tumor Fraction Estimation Performance 47

5 Discussion 49

6 Future Work 51

References 52



AARHUS UNIVERSITY

Thesis Title:

Deep Self-Supervised Modeling of cfDNA Fragmentomics for Tumor Detection

Thesis Period:

Spring Semester 2026

Author:

Ruiqi Zhang

Supervisor:

Jakob Skou Pedersen

Date of Completion:

June 15, 2026

Master's of Science in Bioinformatics

Bioinformatics Research Center (BiRC)

Aarhus University Denmark

In collaboration with:

Department of Molecular Medicine (MOMA)

Aarhus University Hospital, Denmark

Acknowledgements

First, I would like to extend my sincere gratitude to my supervisor Jakob Skou Pedersen, for providing me with the opportunity to start this thesis and work on this dataset, and for his critical and inspiring discussions with me throughout. I am also greatly thankful to Christian, Mathilde and Camous, for their constructive feedback and essential suggestions that enhanced the quality of my work.

My work is also credited to my classmate Mark with whom I had many meaningful and intuitive discussions. My appreciation goes to all the researchers and staff at the Department of Molecular Medicine and Bioinformatics Research Centre. The supportive environment and available resources have been invaluable to me.

Then, my deep thanks to all who inspired and encouraged me during this project. I would like to extend my innermost thanks to my family for their unconditional support and encouragement. To my parents, thanks is not enough for their unconditional love and for trusting me. I am also grateful to my friends here in Aarhus for their support and for keeping me updated.

Abbreviations

cfDNA	cell-free DNA.
cfRNA	cell-free RNA.
ctDNA	circulating tumor DNA.
mRNA	messenger RNA.
PBMC	Peripheral Blood Mononuclear Cell.
NCBI	National Center for Biotechnology Information
RNA-seq	RNA-sequencing.
SGD	Stochastic gradient descent.
TSS	Transcription Start Site.
WGS	Whole Genome Sequencing.
TCGA	The Cancer Genome Atlas.
CRC	Colorectal cancer.
AUC-ROC	Area Under the Receiver Operating Characteristic.
MAE	Masked Autoencoders.
NB	negative binomial.
MSE	Mean Squared Error.
MIL	multi-instance learning.
SSL	self-supervised learning.

Abstract

Cell-free DNA (cfDNA) fragmentomics provides a non-invasive means of studying chromatin organization and cancer-associated biological processes. However, the raw cfDNA fragments data are sparse and high-dimensional, making downstream tasks challenging. Meanwhile, fragmentation patterns surrounding transcription start sites have emerged as a promising source of information for inferring gene expression and chromatin activity. However, the extent to which these signals can be leveraged to learn sample-level representations for downstream tasks, such as cancer detection and tumor fraction estimation, is not yet well investigated.

In this thesis, a self-supervised learning framework inspired by a masked autoencoder was developed to learn representations of cfDNA fragmentomics features surrounding transcription start sites (TSSs) from whole-genome sequencing (WGS) data of a local colorectal cancer cohort. This thesis investigates how well the embedding learned is biologically informative, and can be leveraged for sample-level cancer detection and accurate tumor fraction estimation.

Results

The learned TSS embeddings exhibited clear associations with gene expression levels despite being trained solely on cfDNA fragmentomics features. Visualization of the embedding space revealed progressive transitions across gene expression deciles.

For downstream prediction tasks, sample-level embeddings generated through candidate-gene pooling outperformed embeddings pooled across all genes, with pooling top 50 candidate genes achieving the best tumor classification performance (AUC=0.865). Tumor fraction estimation achieved a Spearman correlation of 0.585, indicating that this approach is not sufficient to achieve accurate tumor fraction estimation.