



Developing and Validating a Novel cfDNA Fragmentation Metric via Deep Learning

Mark Hegedus

Master's Thesis

June 15th, 2026

Supervisor: Søren Besenbacher



Abstract

This thesis focuses on the challenge of early cancer detection using cell-free DNA (cfDNA) fragmentomics, with an emphasis on chromatin accessibility. Low-coverage WGS (lcWGS) data, and 16 cell-type-specific sets of DNase I hypersensitive sites are used. Fragments around these tissue-specific sets of sites are aggregated into a matrix, where fragment length and relative position to the DHS site serve as indices. Therefore, only fragments near active regions of the DNA are studied. This work introduces and benchmarks three novel deep-learning-based fragmentation metrics across the following architectures: multilayer perceptrons (MLP), convolutional neural networks (CNN), and variational autoencoders (VAE). The models are trained and evaluated on independent cohorts and compared to five previously published accessibility scores regarding tumor classification. Furthermore, the study not only optimizes model hyperparameters but also investigates methodological factors, such as GC bias correction and aggregation strategies. Results show that optimized VAE-based models outperform human-derived accessibility scores in an across-cohort setup, where two independent datasets are used for training and evaluation separately. This result highlights the robustness to the domain shift effect and scalability with increasing training data. The findings provide a framework for fragmentomics-based, non-invasive cancer detection, highlighting the importance of deep learning in cfDNA analysis, capturing relevant signals for early tumor classification.

Contents

Introduction	1
Biological background	1
Cancer as a disease	1
DNA fragments as biomarkers	2
Fragmentomics	4
DNase I hypersensitive sites (DHSs)	5
Deep learning	7
Introduction	7
Genomics usage	11
Prior work	14
Research objectives	17
Methods	19
Data	19
Synthetic negative Lymphoid DHS	20
Preprocessing DHS files	21
Preprocessing cfDNA fragments	21
Initial matrix construction	21
Trimming margins	22
Coverage normalization	23
Rebinning matrices	23
GC bias correction	25
Accessibility scores	25
Long Windowed Protection Score (L-WPS)	25
Integrated Fragmentation Score (IFS)	25
Promoter Fragmentation Entropy (PFE)	26
Orientation-aware cfDNA Fragmentation (OCF)	27
Fragment Dispersion Index (FDI)	28
Model architectures	29
CNN and MLP architectures	29
VAE architecture	31
Training	32
Downstream analysis	33
Results	35
Comparison of accessibility scores with and without GC bias correction	35
Benchmark results	37

Structure of VAE latent representations	39
Optimized VAE configurations.....	41
Discussion	44
Interpretation of the results.....	44
GC bias correction effect on accessibility scores	44
Benchmark interpretations.....	45
VAE results	46
Comparison to related works.....	48
Limitations and future work	49
Conclusions.....	52
Appendix.....	54
References.....	70