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