E-Py-Genetics

TABLE OF CONTENTS

01

Theory

02

Insights

03

Predictive modelling

04

Future Work

05

Closing Remarks

01

Theory

Epigenetic DNA Modification

01

Epigenetic modifications are relevant changes to the genome that do not alter the base sequence itself.

02

These modifications are, amongst other things, involved in the development of cancer and in evolutionary biology.

03

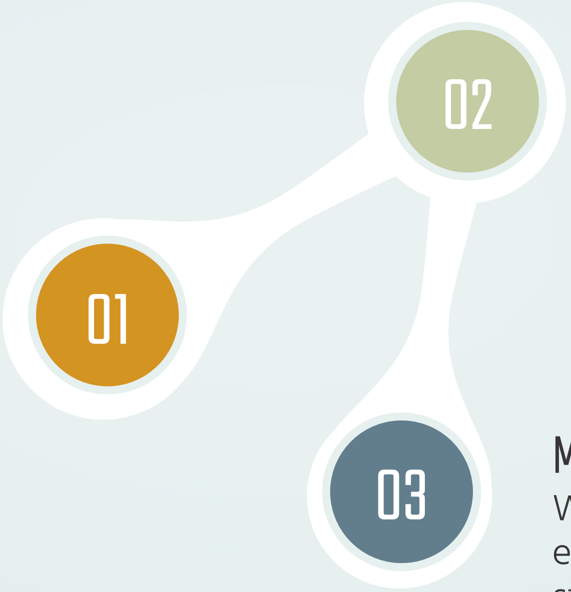
Currently established methods of detection of are comparatively slow and complex.

Nanopore-Sequencing

Data Basis

The utilized Dataset contains multiple reads of 5 different DNA strands of a length of 200 bases, which were either synthetically modified or left unmodified

Analysed parameters



Base Sequence

Which base is on which position in every step of the analysis?

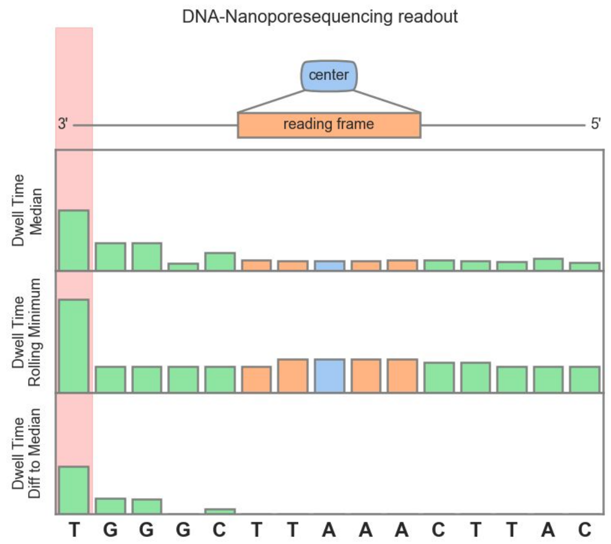
Dwell time

How long does the analysed part of the DNA stay in the pore?

Median Value

What is the value of electrical current at each step of the analysis?

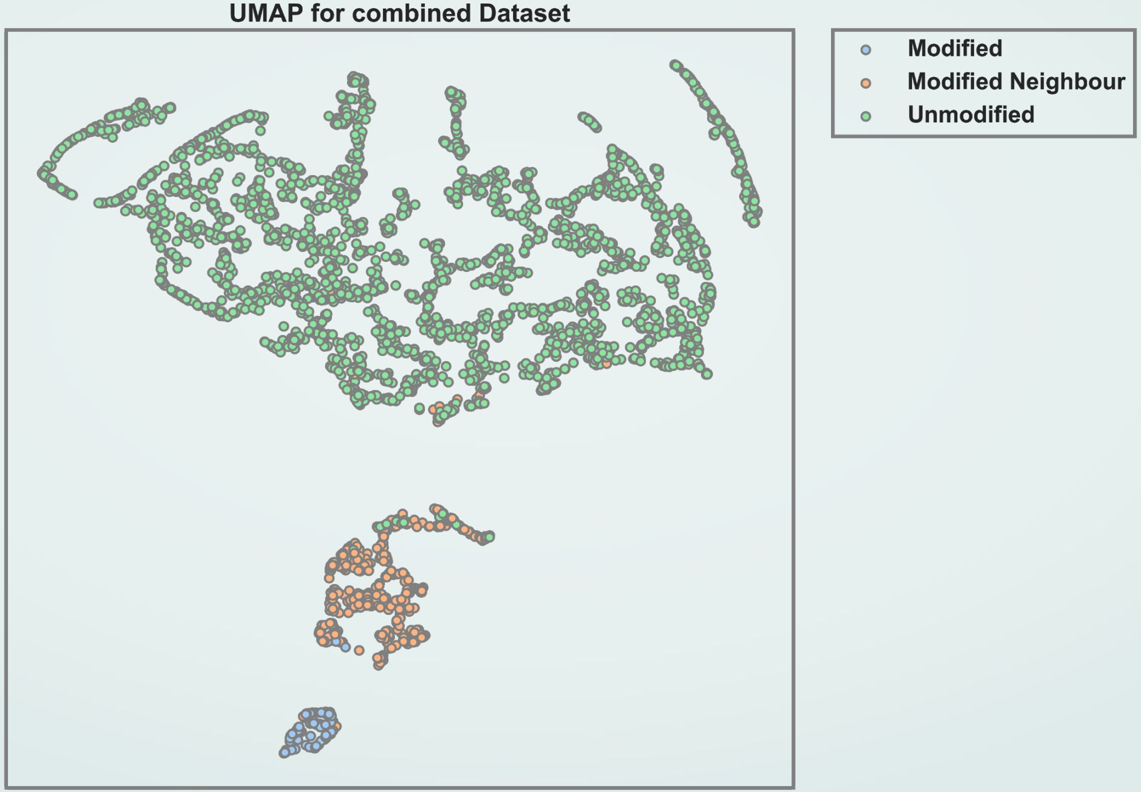
Data Basis


DNA-Nanoporesequencing readout
center
3'
reading frame
5'
Dwell Time Median
Dwell Time Rolling Minimum
Dwell Time Diff to Median
T G G G C T T A A A C T T A C

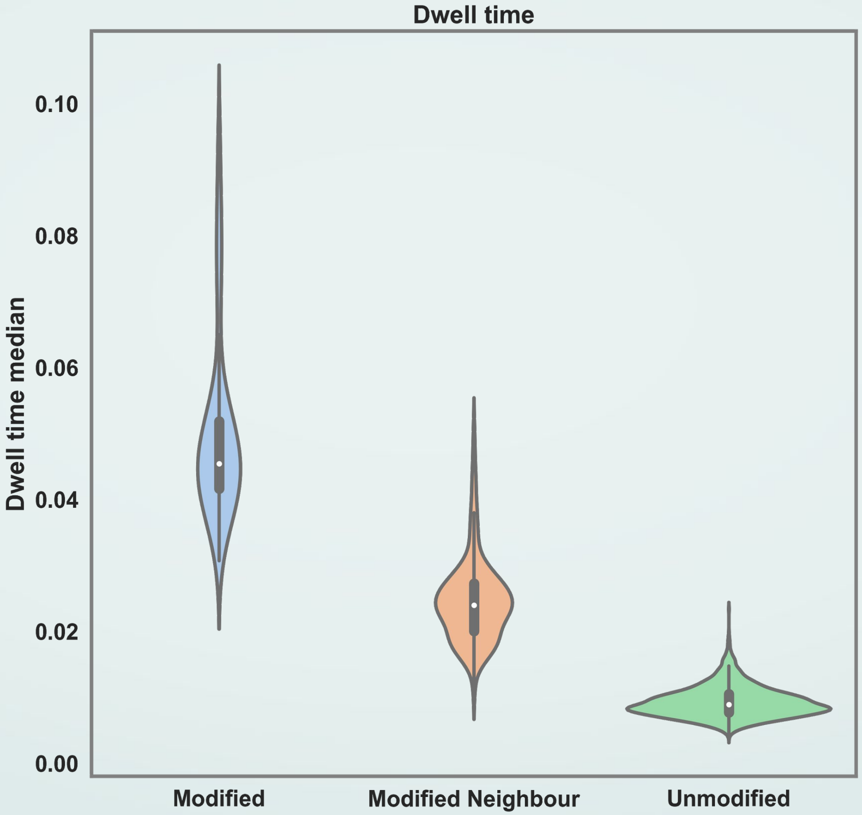
Insights

02

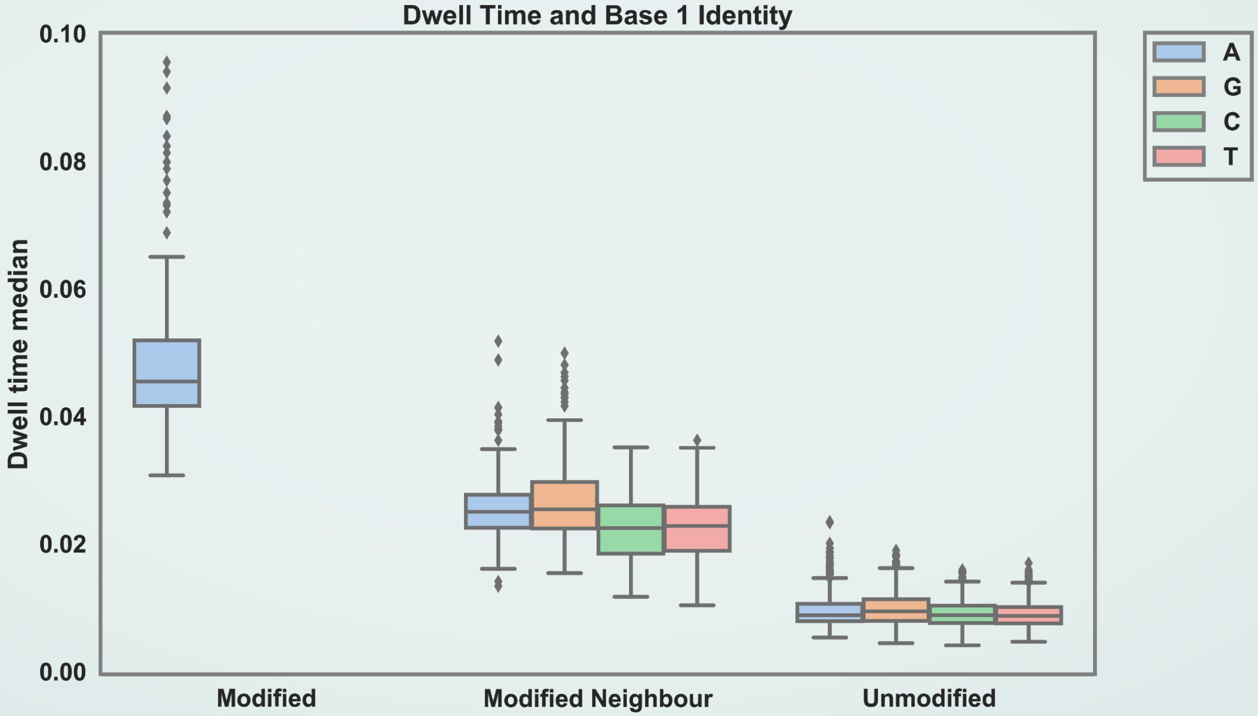
Base modification alters analysis parameters



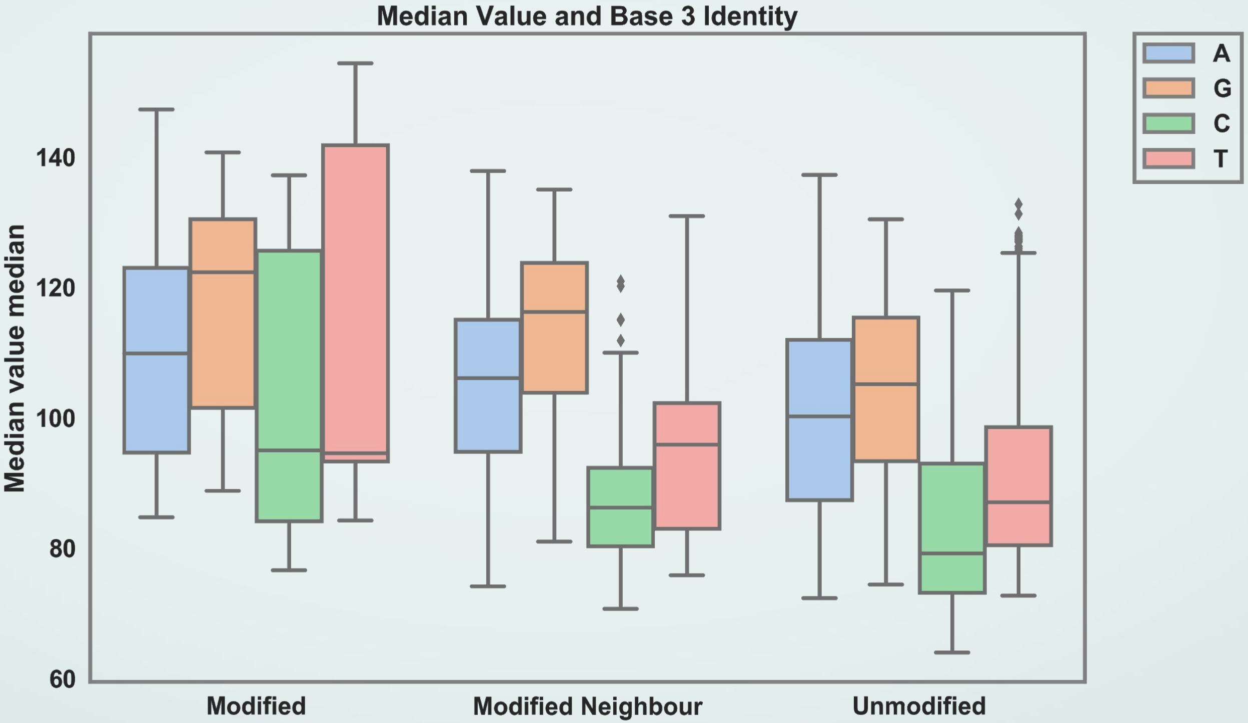
Dwell time increases if a modified base is present



If a modified Base is on position 3, an Adenosin is present on position 1



Base identity has a bigger impact on electric current than base modification



Predictive modelling

03

Ignore Base information

There seems to be information in the base sequence, but in the model it is ignored, because it is more comparable to real world applications

Ignore electric current

Electric current is more dependent on base identity, than on base modification

Train model on two Dataset versions

Version 1: Treat every read as an individual observation

Version 2: Aggregate reads

Feature Engineering

Several new features were calculated from experimental data, to highlight differences between modified and unmodified parts of the DNA

Comparison of ML models

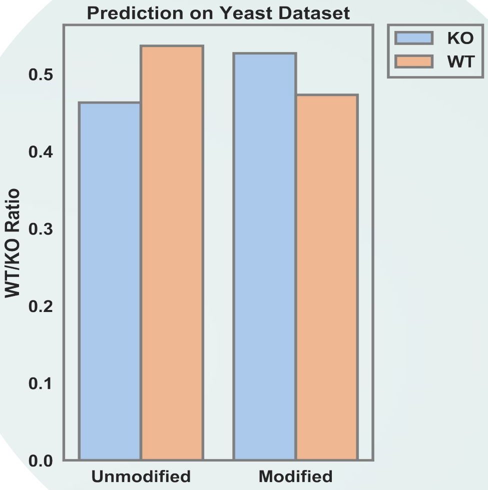
Several different ML models were trained on the dataset, to select the best performing models

Model finalization

The best performing models on the two versions of the dataset were fine-tuned, combined and validated

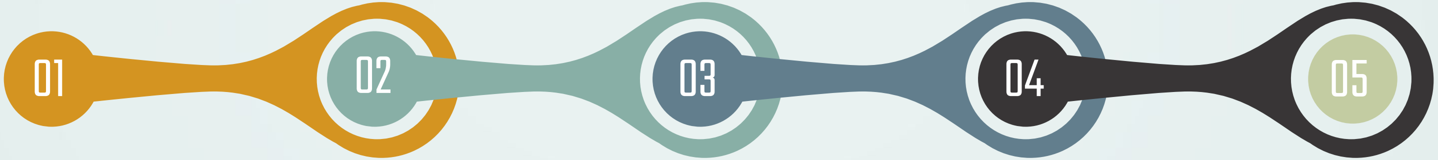
Best Model

- Usage of the state of the art XGBoost algorithm on the aggregated Dataset proved to be the best performing model (F1-Score 96%) in comparison to the unaggregated Dataset (F1-Score 79%)
- Usefulness on different Datasets seems to be limited (Training data was simulated)



04

Future Work



Neural Networks

Improving Recurrent Neural Network architecture

Unsupervised Learning

Implementation of unsupervised outlier detection, to analyse unlabelled datasets

Additional Datasets

Improve ML model using a bigger dataset

Cloud Services

Port ML models to cloud services, to improve modelling times

Deployment

Package prediction models for application in real workflow

05

Closing Remarks

Thank you for Your attention!

- Thank you to neuefische, especially Larissa and Dirk, for the great Bootcamp, that allowed me to learn all the DataScience skills demonstrated in this capstone project
- Thank you to Evotec, and especially Benedikt, for the cooperation in executing this capstone project
- And a big thank you to the DataScience cohort, who made it a pleasure to endure this Bootcamp with

https://github.com/Karsten-Yan/ky-nf-capstone

RESOURCES

Dataset

- https://github.com/tleonardi/nanocompore/

Opening Gif

- https://nanoporetech.com/how-it-works

Tech Stack

- Python
- Pandas
- Scikit_learn
- Tensorflow
- Keras
- Ensemble Methods (XGBoost, ADABoost, Stacking)
- Matplotlib
- Seaborn

THANKS

Do you have any questions?

youremail@freepik.com
+91 620 421 838
yourcompany.com

Please keep this slide for attribution