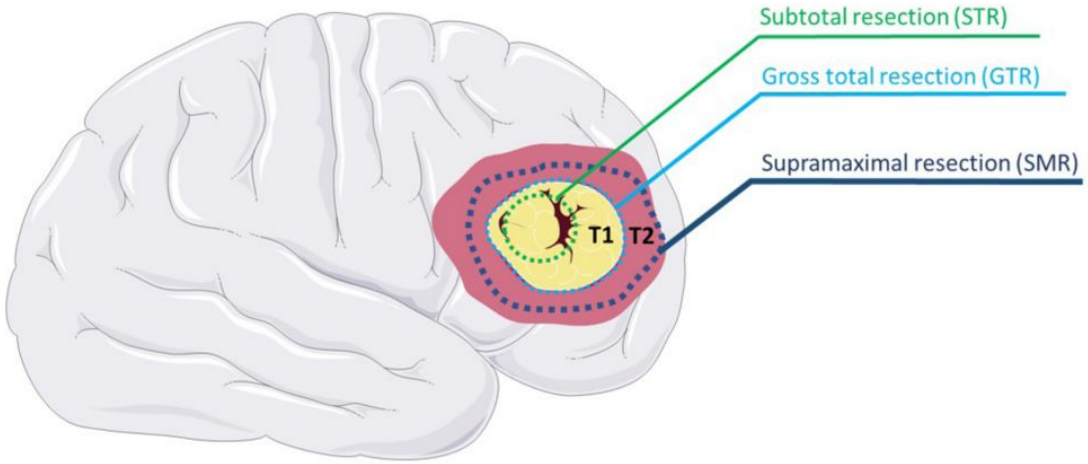
MAX PLANCK INSTITUTE
FOR MOLECULAR GENETICS

MethyLYZR: Live brain tumor diagnosis from sparse epigenomic data

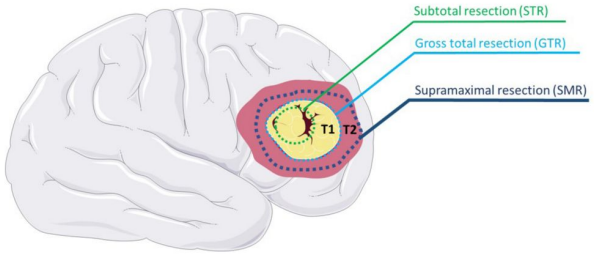
Mara Steiger
Max Planck Institute for Molecular Genetics, Kretzmer Lab

Brain tumor resection



Willman et al., 2023
Badza et al., 2020

Standard practice for CNS tumors



1 Preparation and begin 60 mins

2 Incision 30 mins

3 Resection 60 min

4 End 30 min

5 Standard molecular diagnostics 4-7 days

Classification

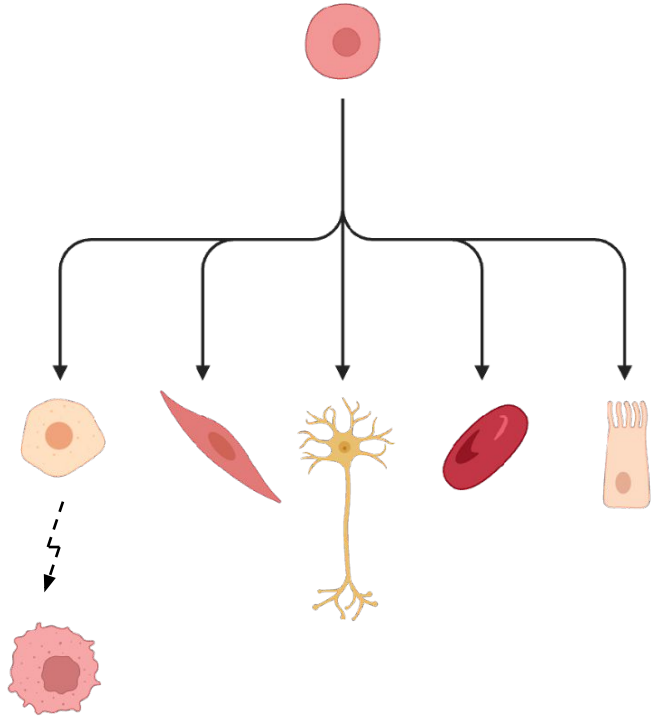
- sequencing
- Immunohistochemistry
- **methylation microarrays**

aim: intra-operative DNA methylation-based classification

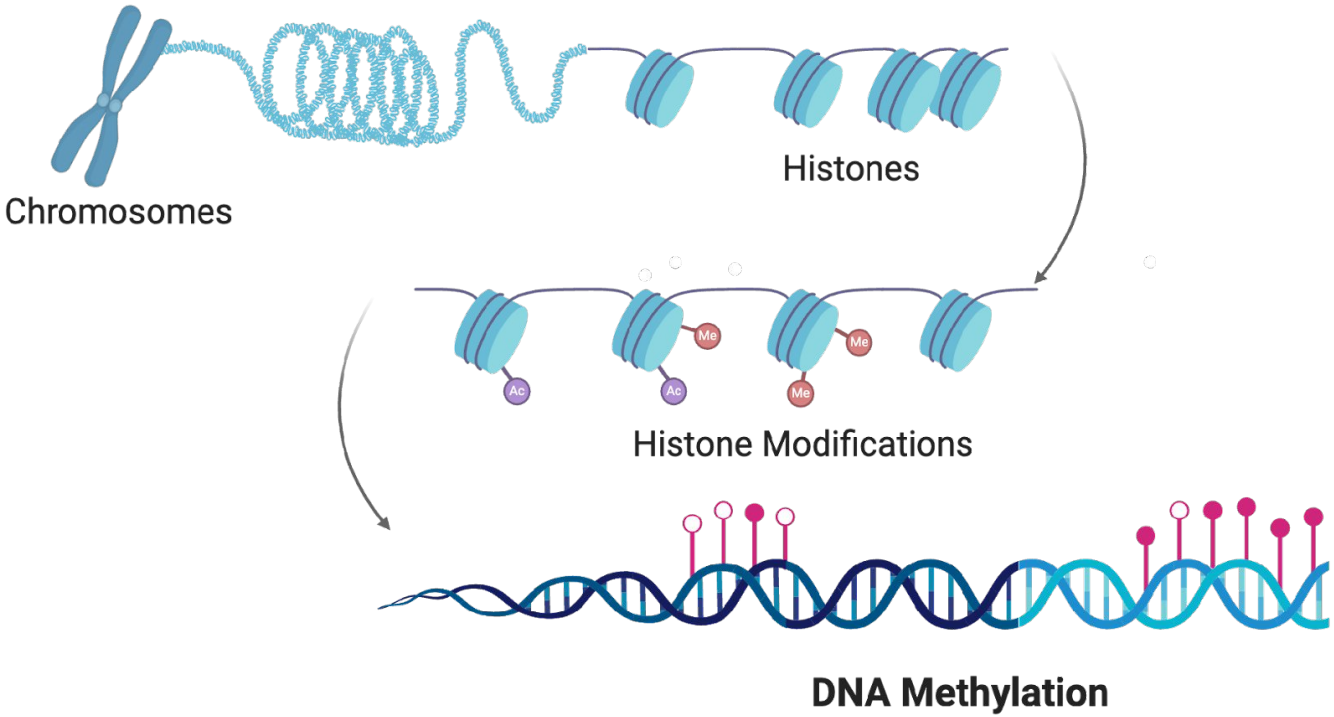
Senders et al., 2018
Sturm et al., 2023

DNA Methylation as biomarker

One genotype
– multiple phenotypes –

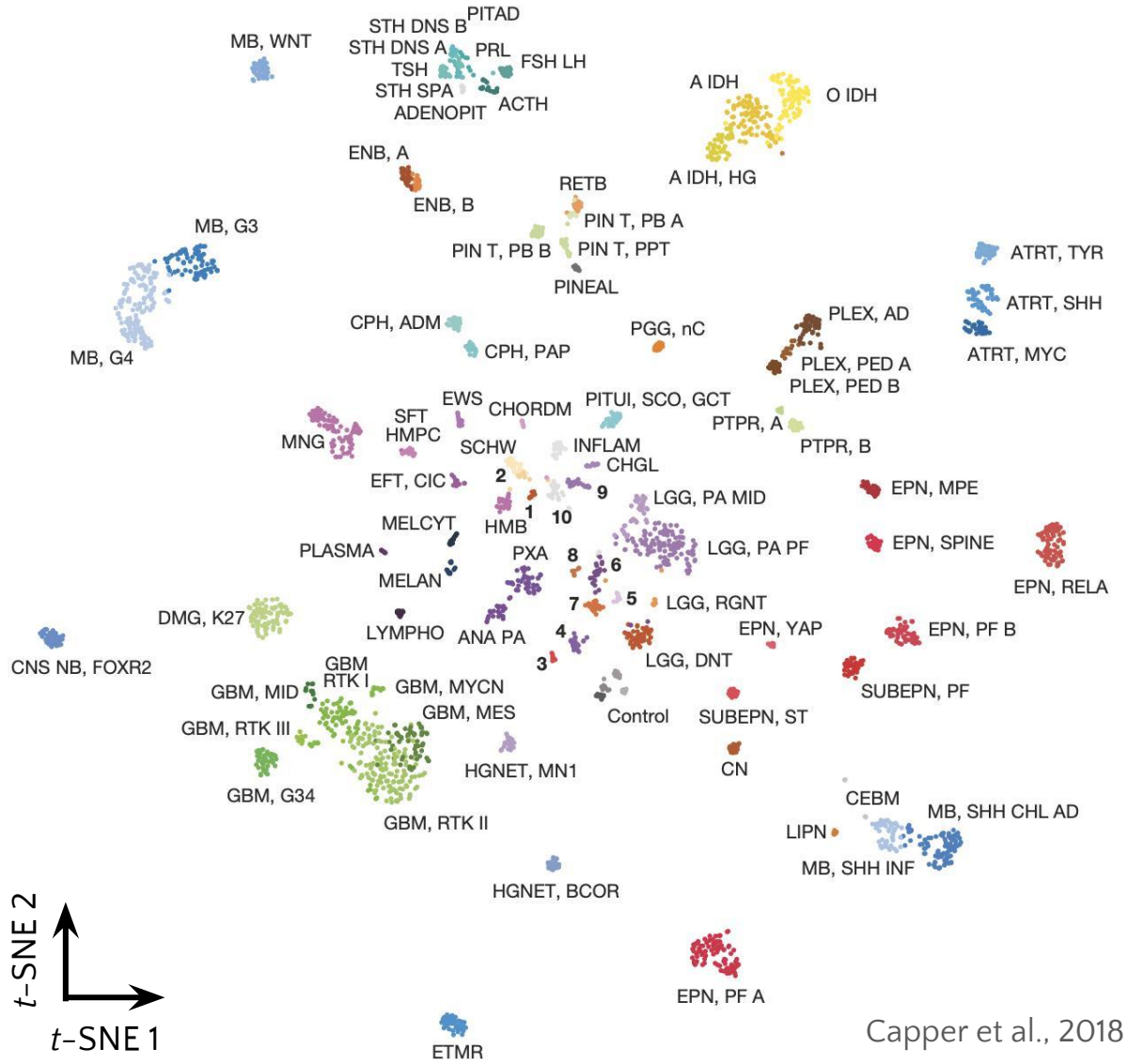


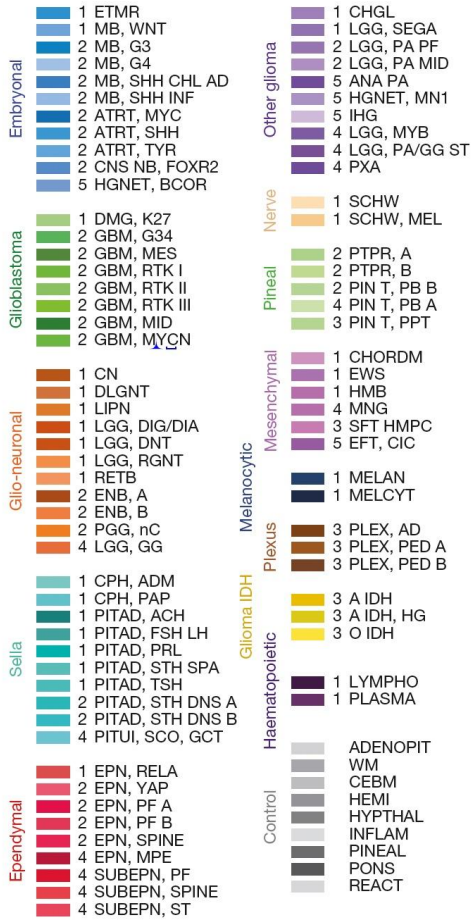
Epigenome
– 2nd layer of information –



Comprehensive training dataset

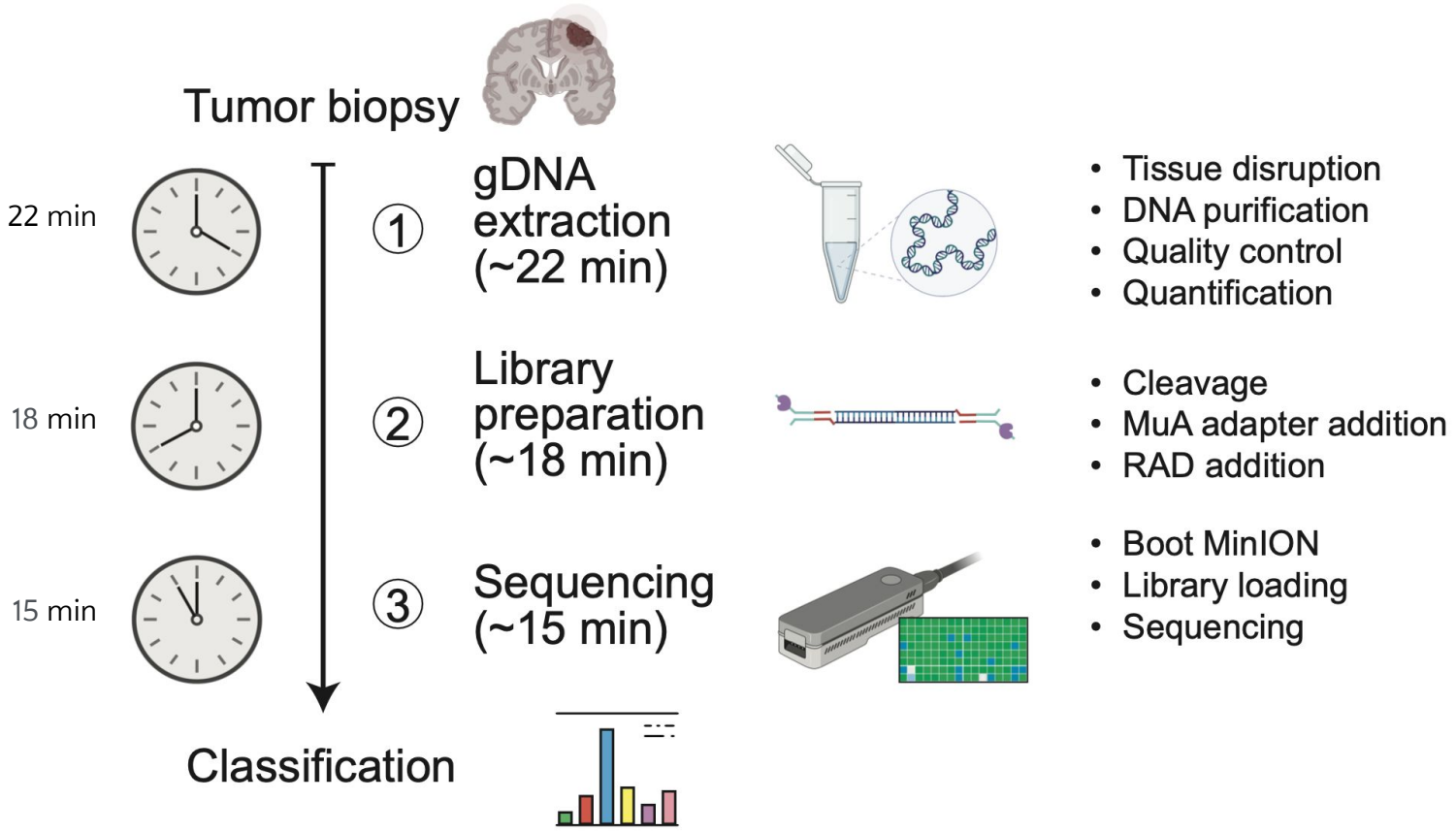
More than 2,800 samples covering 91 tumor classes





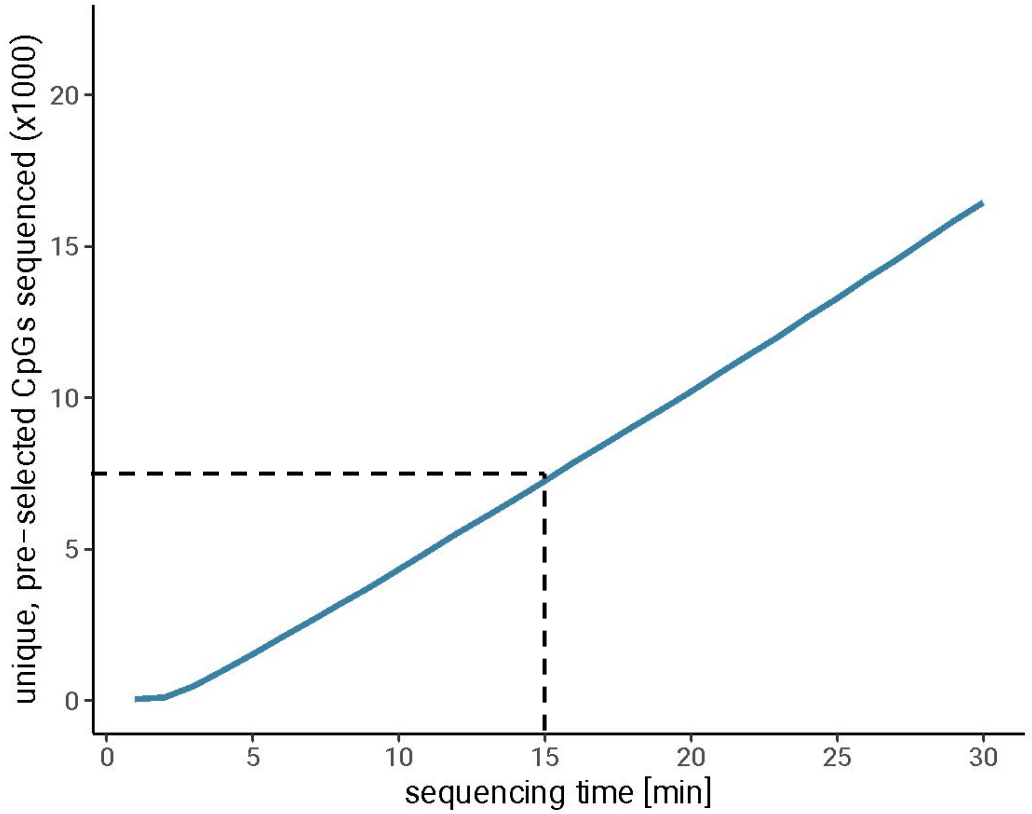
Capper et al., 2018

Intra-operative classification approach



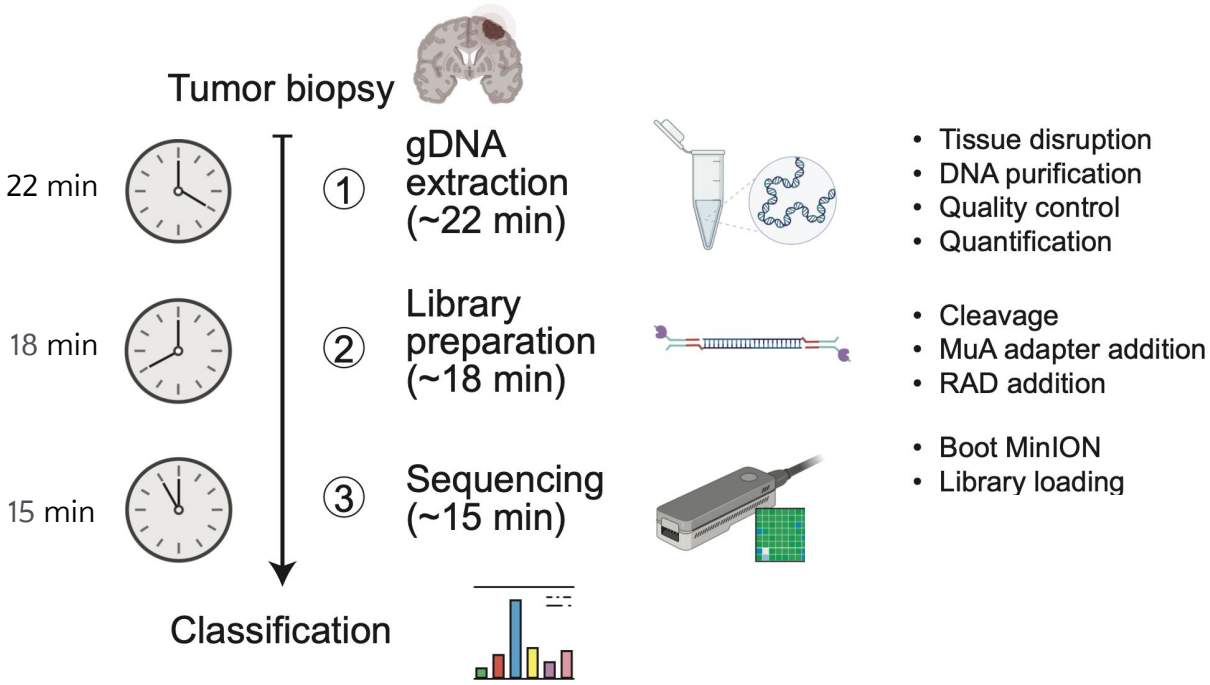
Intra-operative classification approach

Throughput (CpGs covered by Illumina 450k BeadChip)



~7,500 CpGs in 15 minutes

98% missing values



Naïve Bayes

Bayes' Theorem

likelihood

class prior

$$P(C_j|X) = \frac{P(X|C_j)P(C_j)}{P(X)}$$

posterior

predictor prior

class C_j

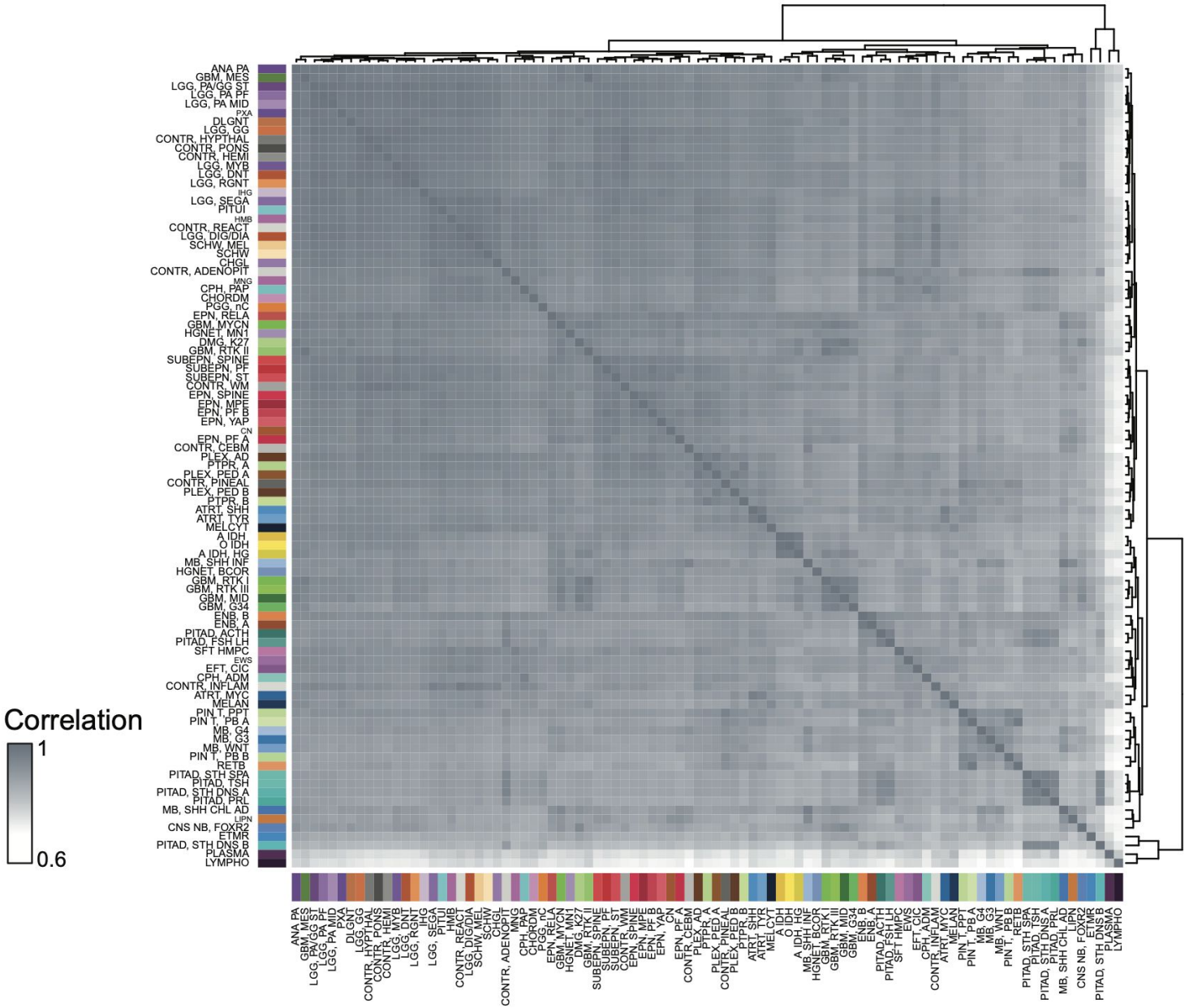
observation X

Independence assumption

$$P(X|C_j) = P(x_1, \dots, x_p|C_j) = \prod_{i=1}^{p} P(x_i|C_j)$$

$$X = \{x_1, x_2, \dots, x_p\}$$

Methylation profiles are highly correlated



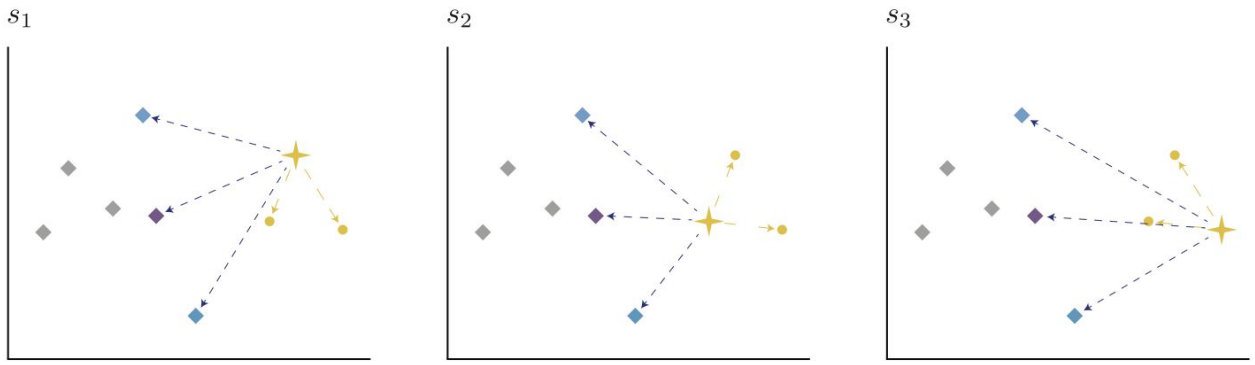
Feature weighting to discern informative and non-informative features

ReliefF-based feature weights

Case I: ▲ inter-class distance ▼ intra-class distance ⇒ $\omega_{i,j} > 0$

$$\omega_{i,j} = \sum_{s \in C_j} \{\text{mean distance to misses} - \text{mean distance to hits}\}$$

for each instance in class C_j

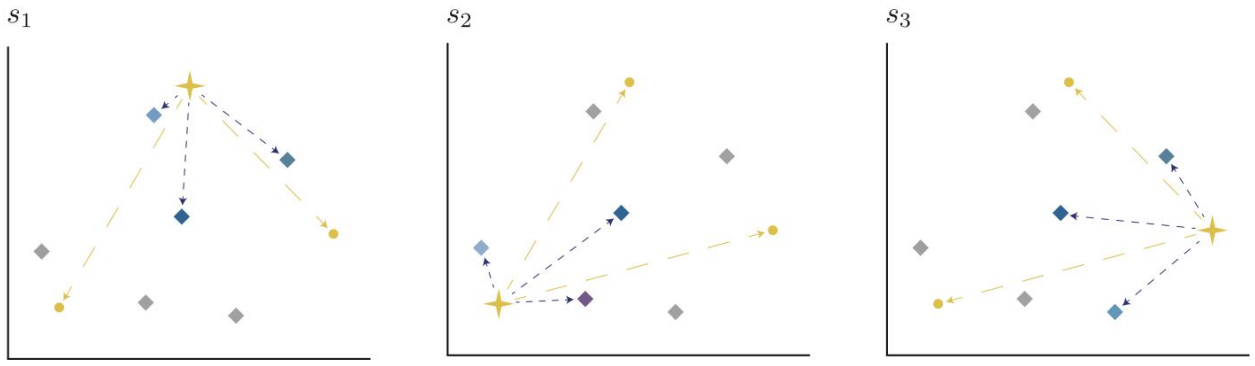


$k = 3$ misses

◇ class centroid
◆ miss
◆ not considered
✦ target instance
● non-target instance
→ distance
⇢ ... to miss
⇢ ... to hit

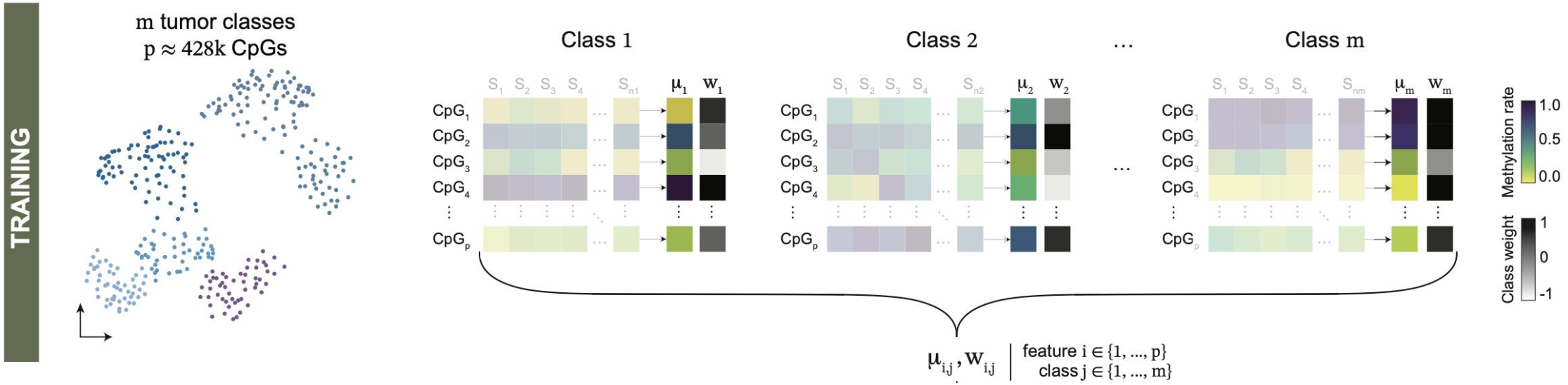
Case II: ▼ inter-class distance ▲ intra-class distance ⇒ $\omega_{i,j} < 0$

for each instance in class C_j



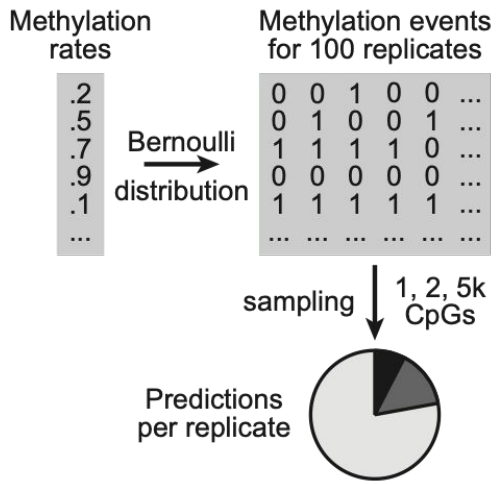
Kira and Rendell, 1992
Kononenko, 2005

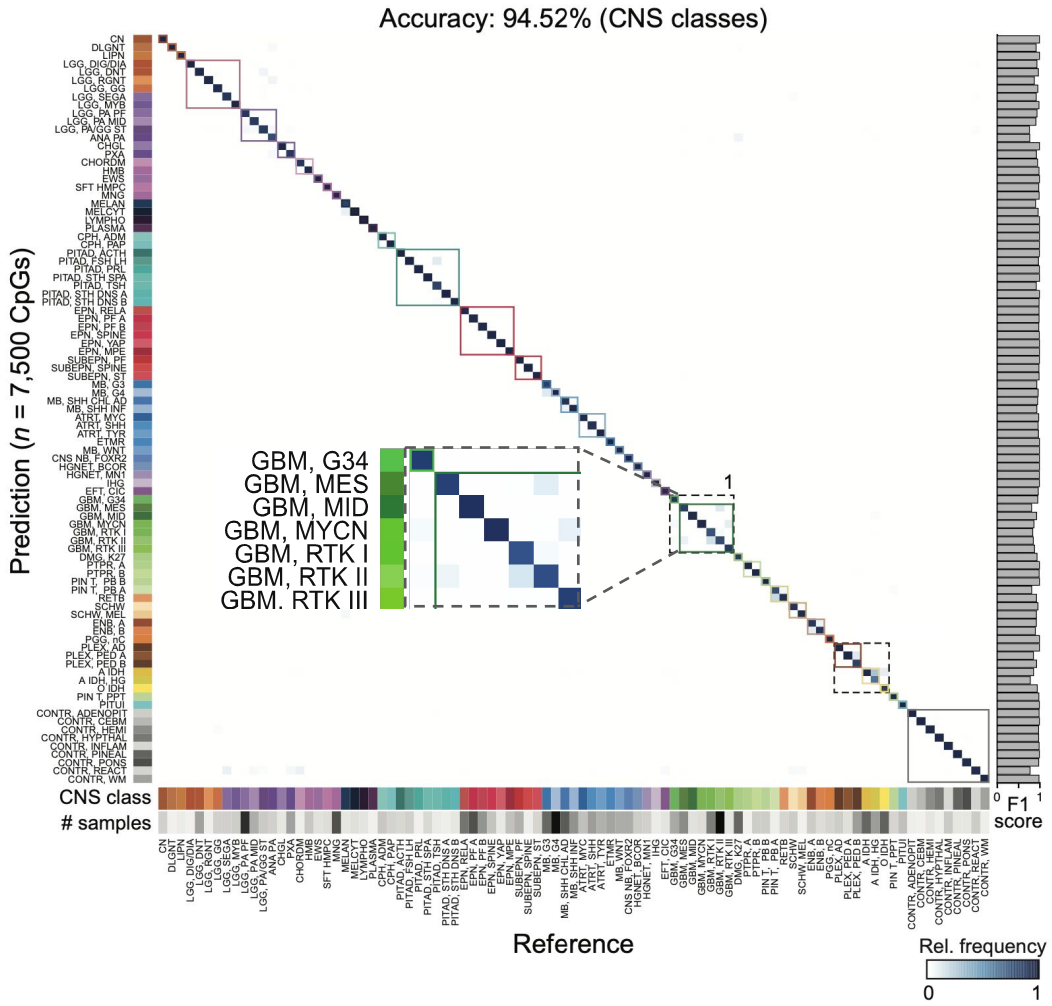
MethyLYZR Framework



Evaluation – Synthetic Data

Using 450k data to simulate low-coverage Nanopore sequencing

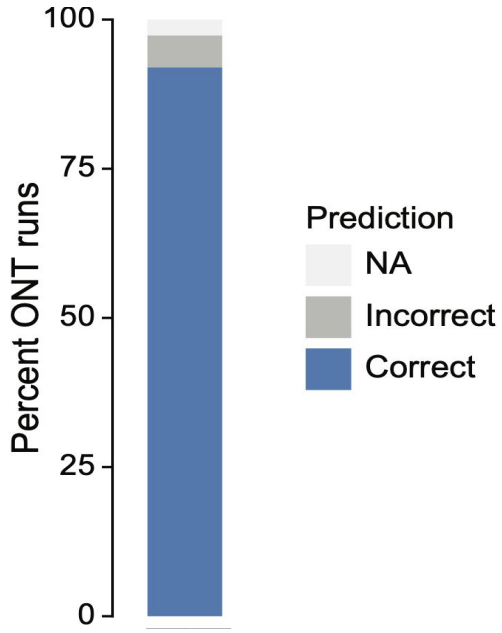


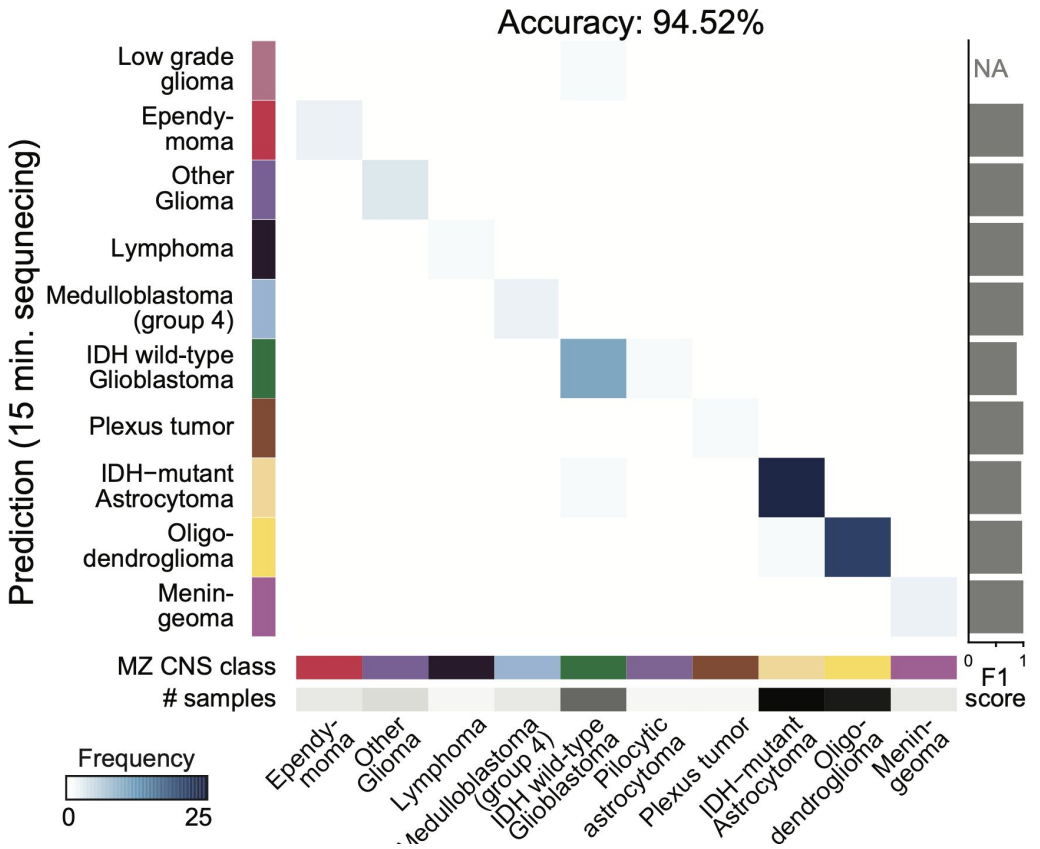


Broad level accuracy: 97.72%

Evaluation – Nanopore data

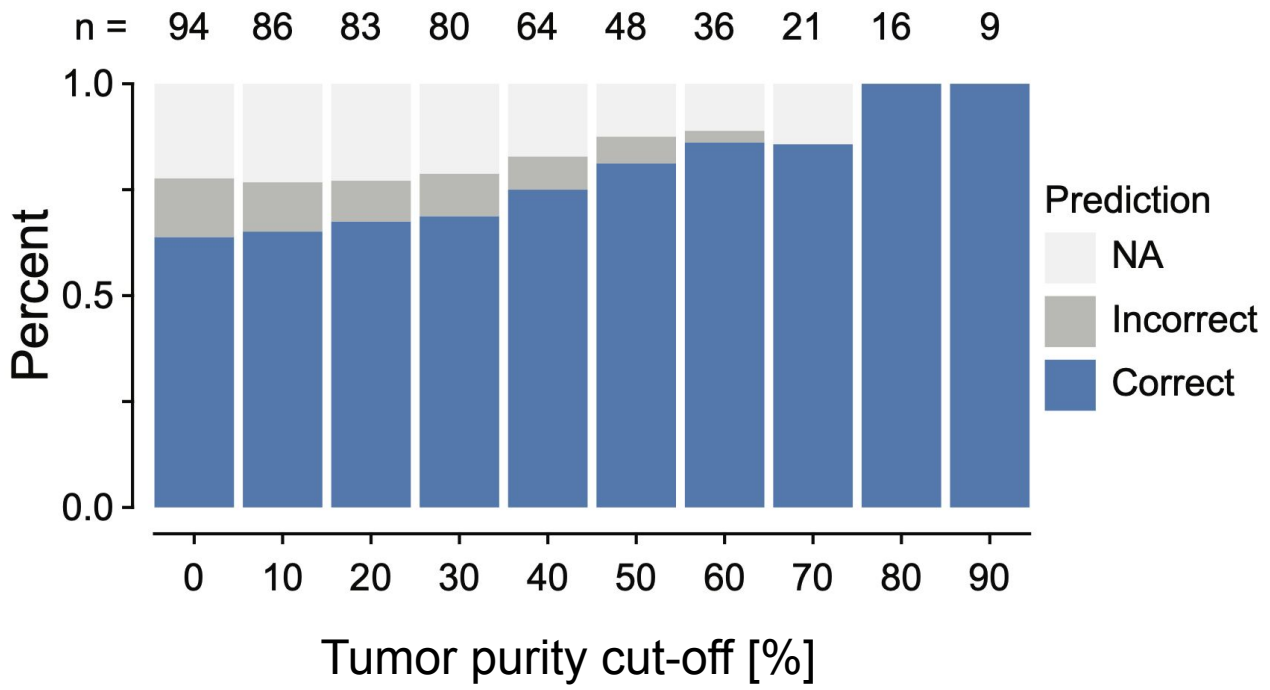
$n = 75$ Nanopore-sequenced samples





Feasibility – Tumor Purity

$n = 94$ external Nanopore-sequenced samples



Acknowledgments

Helene Kretzmer

Franz-Josef Müller

Alena van Bömmel

Björn Brändl

Carolin Kubelt

Michael Synowitz

Christian Rohrandt

Bernhard Schuldt

Gaojianyong Wang

Romulas Smičius

Maximilian Evers

Stephen Yip

Ole Ammerpohl

UKSH UNIVERSITÄTSKLINIKUM Schleswig-Holstein

FACHHOCHSCHULE KIEL University of Applied Sciences

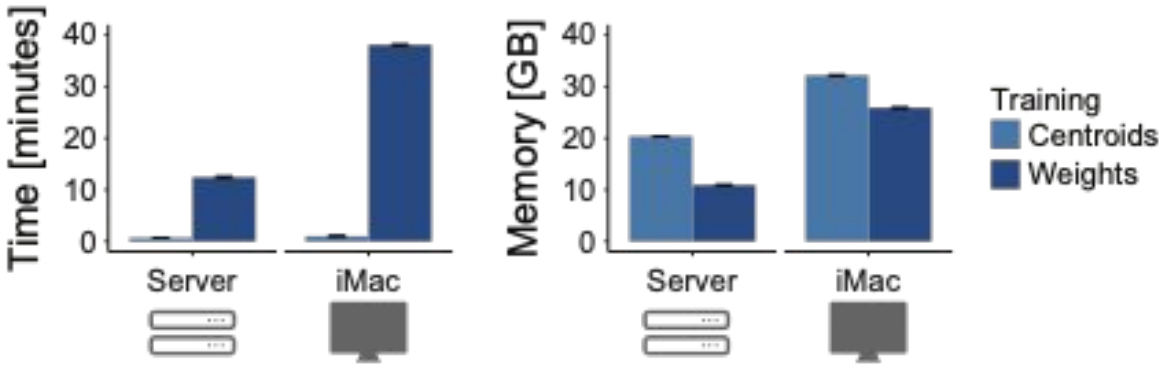
MAX PLANCK INSTITUTE FOR MOLECULAR GENETICS

altona DIAGNOSTICS

Federal Ministry of Education and Research

Funding: IntraEpiGliom
FKZ 13GW0347

Training Naïve Bayes model



Complexity

- p: number of features, i.e., CpGs
- m: number of classes
- n_j: number of samples in class C

  $\rightarrow$ such that $N = \sum_j n_j$ is the number of all samples

centroids

$$\mathcal{O}(p \cdot N)$$

weights

$$\mathcal{O}(m \cdot N \cdot p) + \mathcal{O}(p \cdot (n_1^2 + n_2^2 + \ldots + n_m^2))$$

Model Scalability

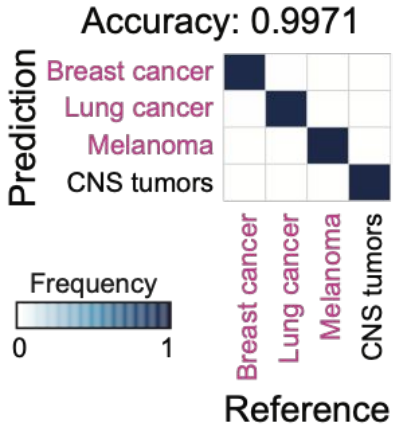
Metastases

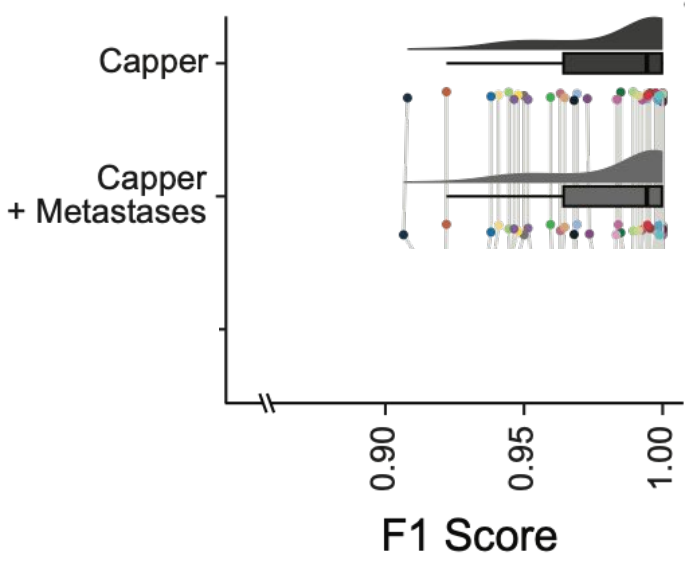
- Breast cancer (n = 30)
- Lung cancer (n = 18)
- Melanoma (n = 37)

Sarcomas

- m = 65 classes
- n = 1077 samples

Synthetic data



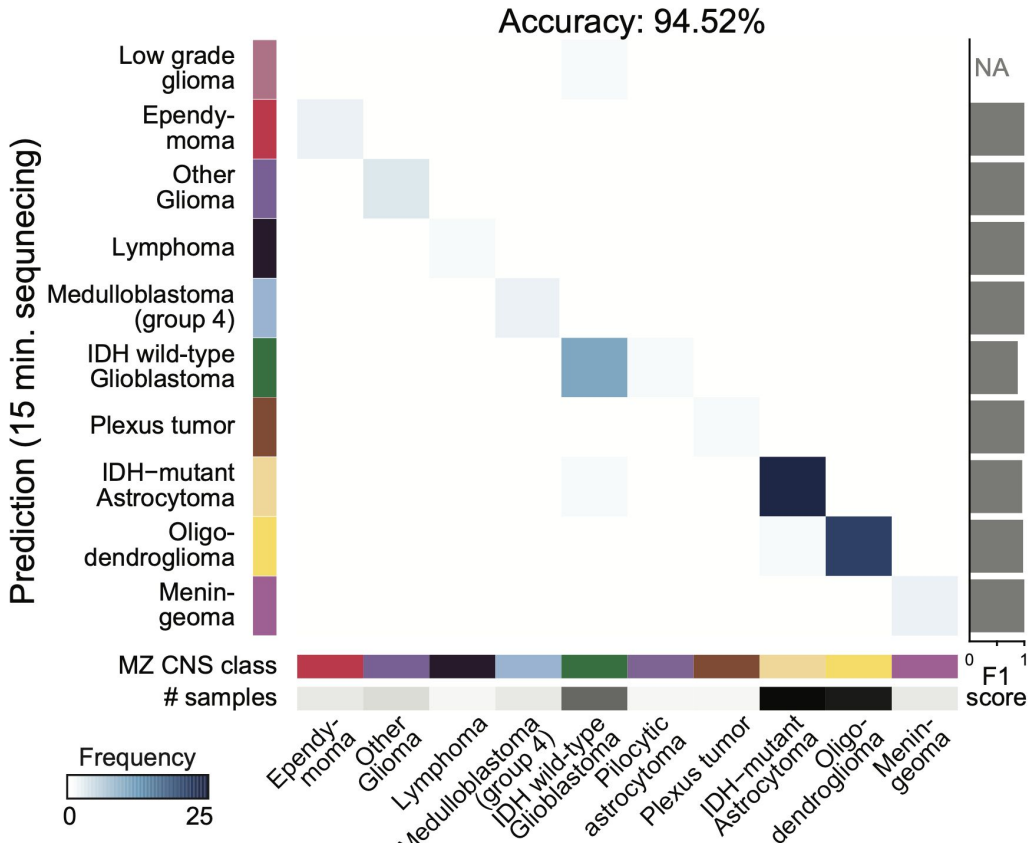


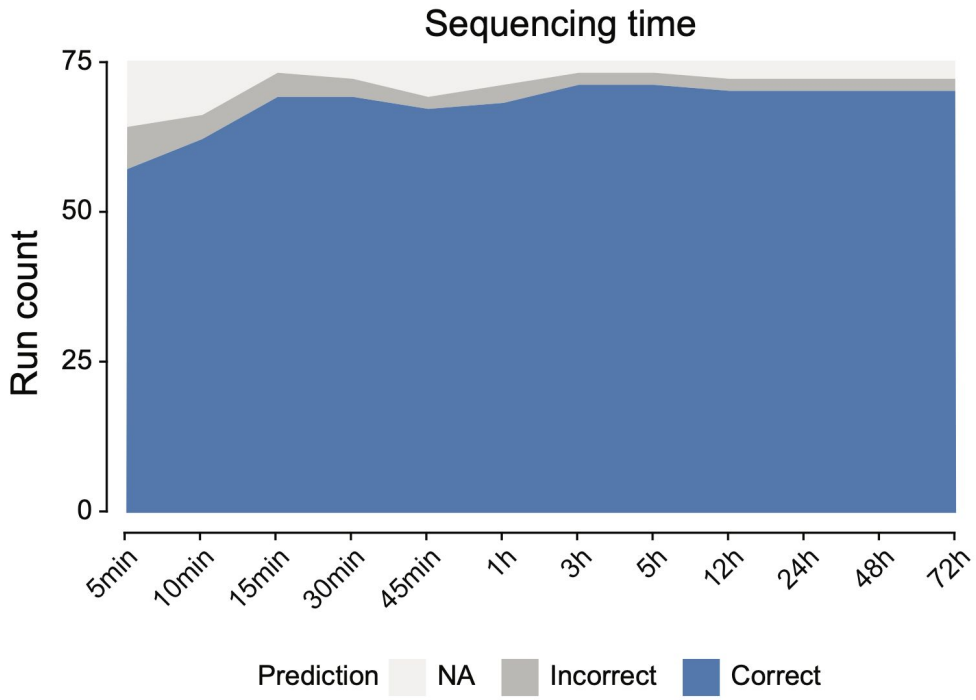
core MethyLYZR classes

Salomon et al., Scientific data (2018)
Koelsche et al., Nat. Commun. (2021)

Evaluation – Nanopore data

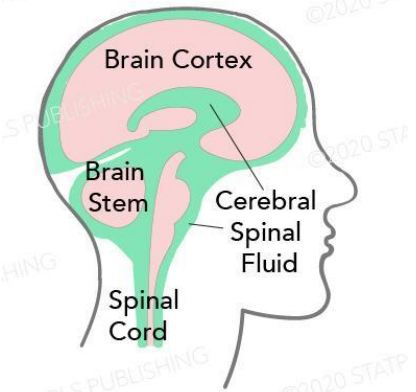
$n = 75$ Nanopore-sequenced samples

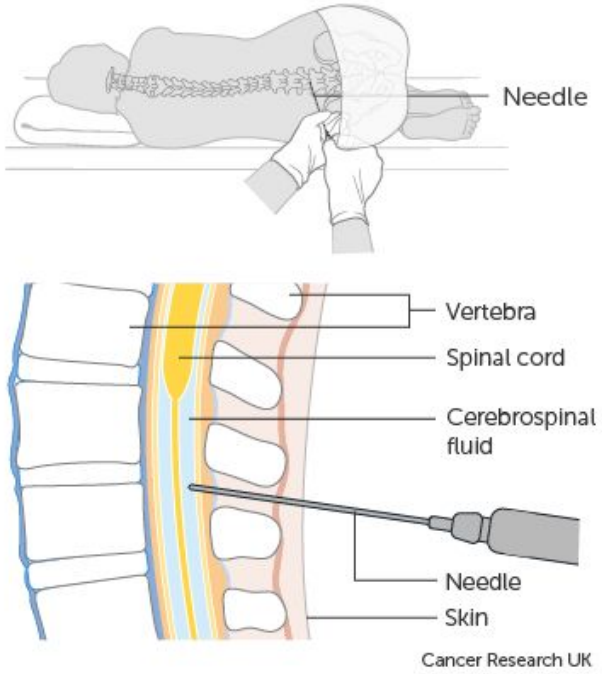




Feasibility – Liquid biopsy

Cerebrospinal Fluid (CSF) from Lumbar Puncture

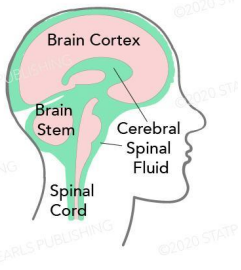


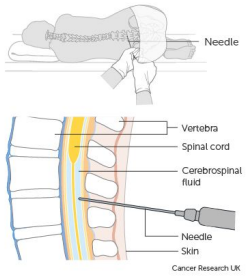


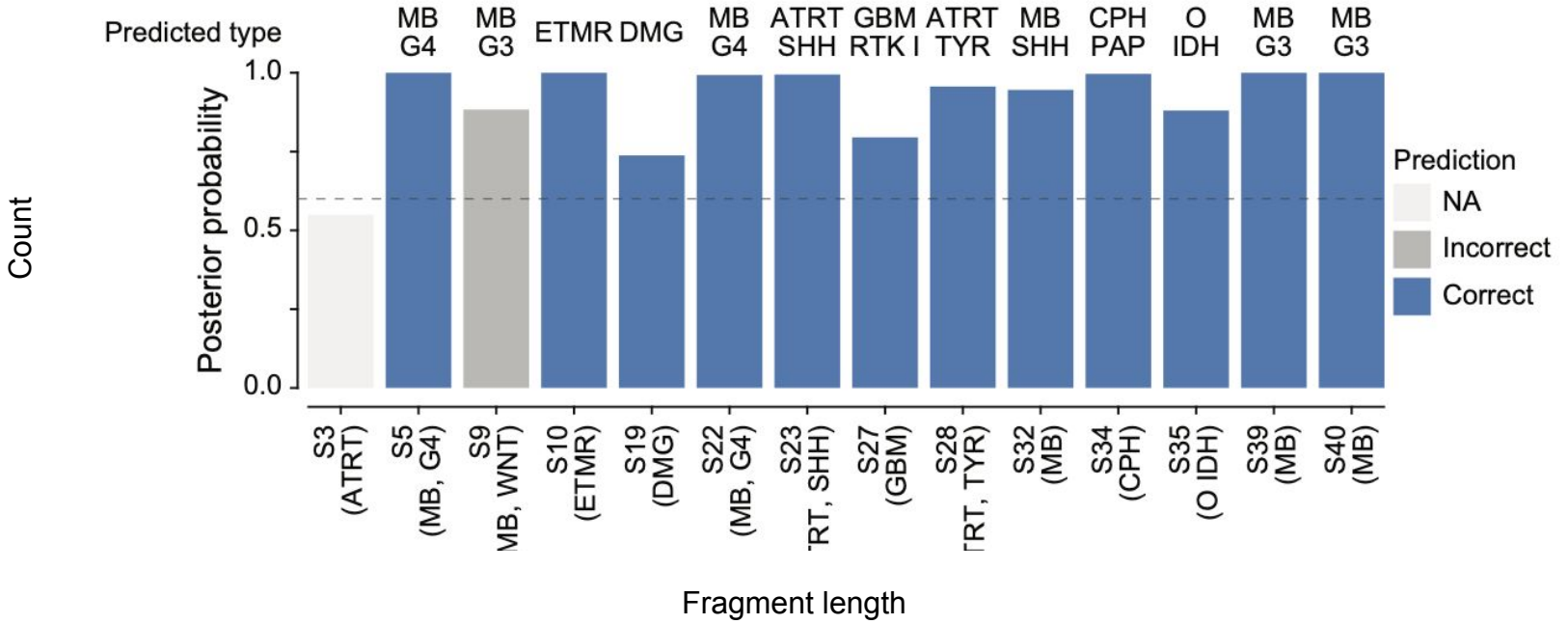
Afflerbach et al., 2023

Feasibility – Liquid biopsy

Cerebrospinal Fluid from Lumbar Puncture







Afflerbach et al., 2023

Evaluation – Nanopore cohort

Comparing Performance to Sturgeon

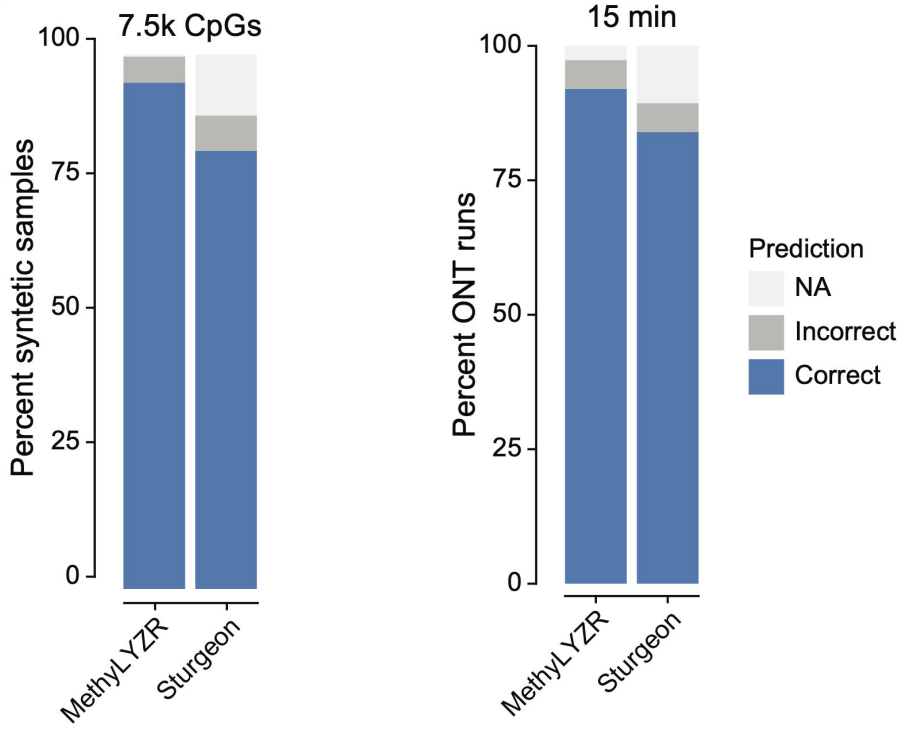
Article

Ultra-fast deep-learned CNS tumour classification during surgery

https://doi.org/10.1038/s41586-023-06615-2
Received: 10 February 2023
Accepted: 6 September 2023

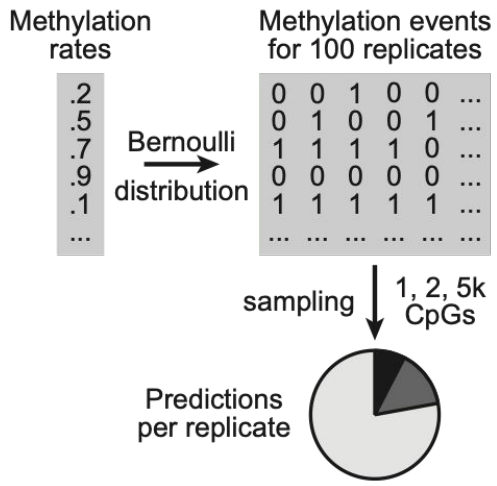
C. Vermeulen[1,2,6], M. Pagès-Gallego[1,2,6], L. Kester[3], M. E. G. Kranendonk[3], P. Wesseling[3,4], N. Verburg[5], P. de Witt Hamer[5], E. J. Kooi[4], L. Dankmeijer[4,5], J. van der Lugt[3], K. van Baarsen[3], E. W. Hoving[3], B. B. J. Tops[3]✉ & J. de Ridder[1,2]✉

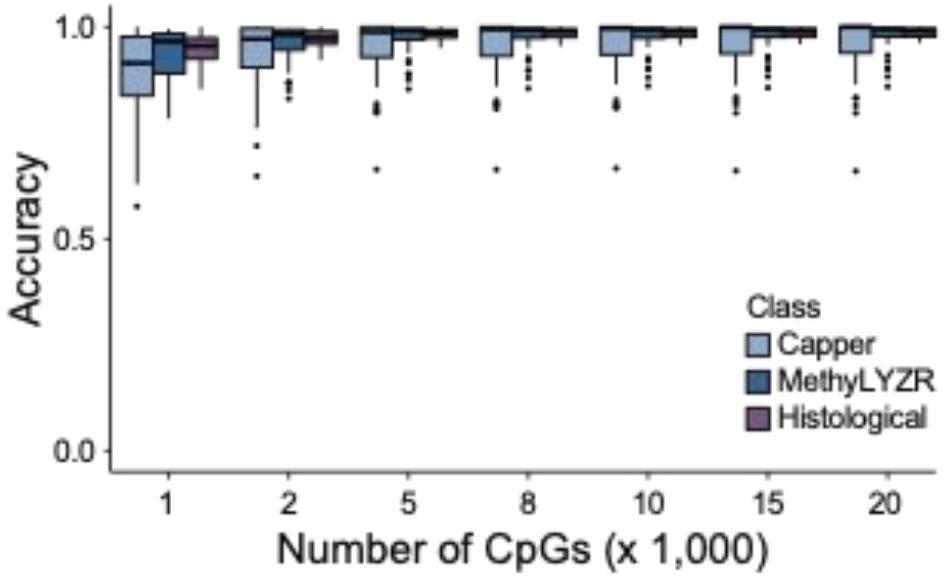
- Data augmentation to train neural network
  - >35 million simulated runs from 2800 arrays
- 18/25 classified samples in under 90 minutes
  - 20–40 minutes of sequencing



Evaluation – Synthetic Data

Using 450k data to simulate low-coverage Nanopore sequencing





Histological classes (m = 8) defined by Capper et al. (2018)

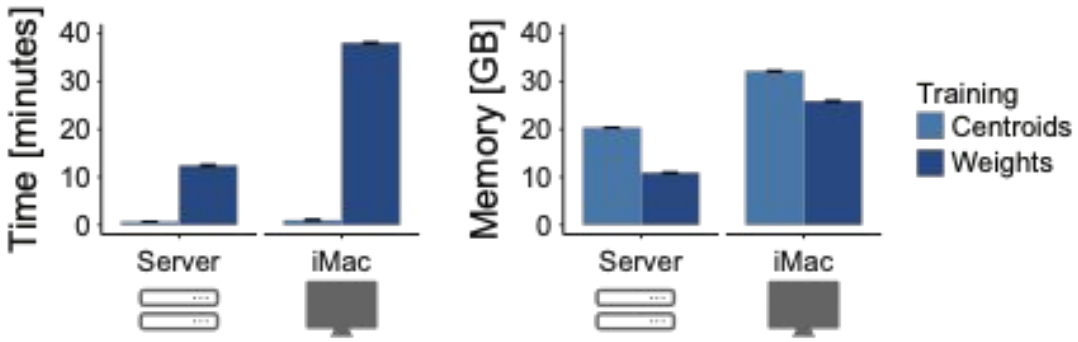
"histologically and biologically closely related tumour classes, the distinction of which is currently without clinical impact"

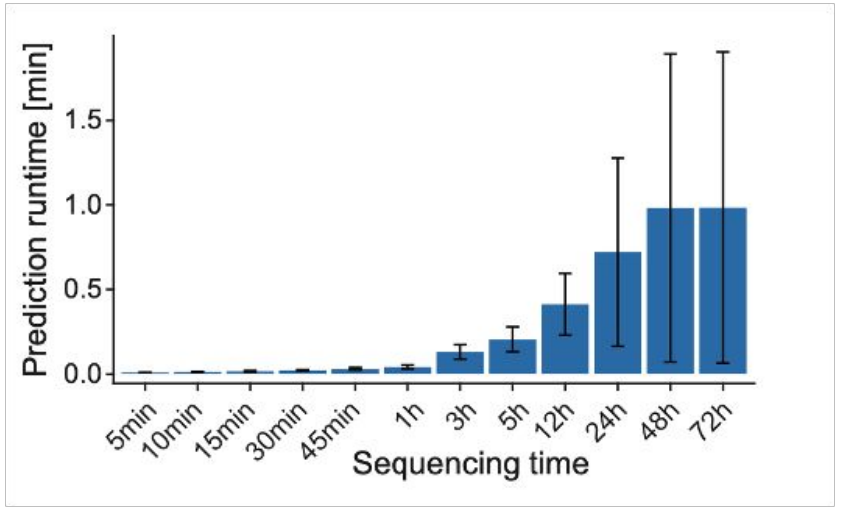
Benchmarking

Training

Prediction

MethyLYZR





Alternative Approaches

- ad hoc RF
  - no pre-training possible
- neural network
  - days – weeks
  - hundreds of GB to TB memory

- ad hoc RF
  - 17-60 min
- neural network
  - few seconds

Djirackor et al., 2021
Vermeulen et al., 2023

RELIEF-based feature weighting

$w_{i,j}$ feature weight for class C_j and feature i

$$w_{i,j} = \sum_{x \in C_j} \left\{ \frac{\sum_{\substack{m \in KNN(x) \\ l(m) \neq C_j}} |x_i - m_i|}{k} - \frac{\sum_{\substack{h \neq x \\ h \in C_j}} |x_i - h_i|}{|C_j| - 1} \right\}$$

, where $KNN(x)$: k-nearest centroids of x,
$l: X \rightarrow C$: maps centroid to class

distance to misses $\boldsymbol{m}$ distance to hits $\boldsymbol{h}$

- $w_{i,j} > 0$: informative feature
- $w_{i,j} < 0$: uninformative feature

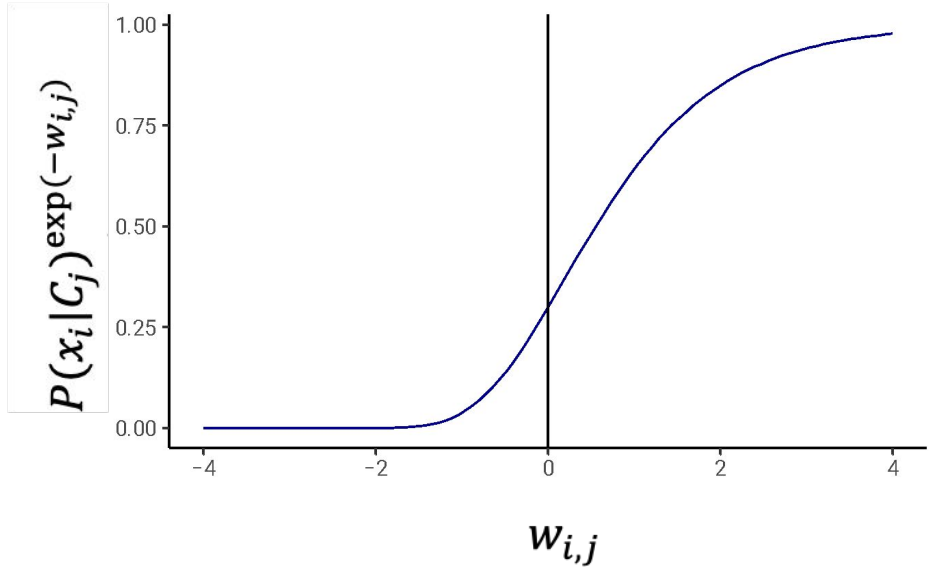
Kira and Rendell, 1992
Kononenko, 2005
Urbanowicz et al., 2018

Feature weighting in Naïve Bayes

proposed by Foo et al., 2021:

$$P(x_i|C_j)^{\exp(-w_{i,j})}$$

Example: $\boldsymbol{P(x_i|C_j) = 0.3}$



Zaidi et al., 2013
Xiang et al., 2015
Jiang et al., 2018