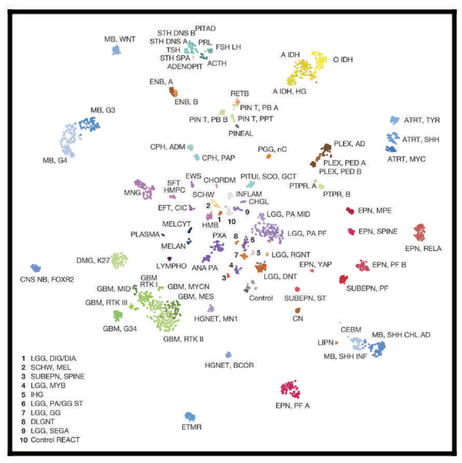
Sturgeon

Ultra-fast deep-learned pediatric CNS tumor classification during surgery

C. Vermeulen[1†], M. Pagès-Gallego[1†], L. Kester[2], M.E.G. Kranendonk[2], P. Wesseling[2,3], J. van der Lugt[2], K. van Baarsen[2], E.W. Hoving[2], B.B.J. Tops[2] & J. de Ridder[1✉]

Introduction

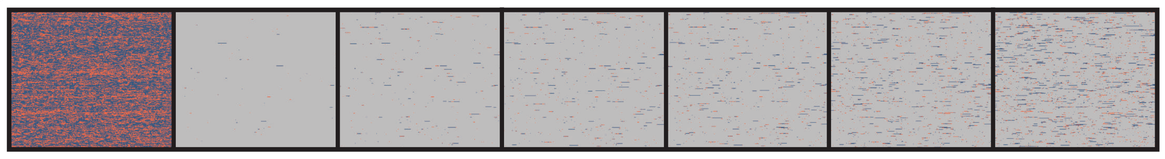
A major challenge in the treatment of central nervous system (CNS) cancer is that taking a biopsy is often not possible. Therefore the precise tumor type is often unknown at the time of surgery, while a choice must be made between radical and conservative resection. Rapid histological analysis is currently the only option to determine the surgical strategy. After surgery, a molecular classification is made to determine the exact cancer type based on the methylome of the tumor. In a number of cases, the molecular diagnosis differs from the histological assessment: this means that either an additional and more radical surgery is needed; or it reveals that the resection was too radical and side-effects of the surgery could have been avoided.



Molecular classification is based on DNA methylation patterns, these are highly distinctive between different types of CNS cancers (see the t-SNE plot above). In routine practice, Illumina Infinium arrays are an essential tool to discern the molecular class of CNS cancer, obtaining methylation signal in hundreds of thousands of CpG sites. A major drawback is that these arrays take several days to process.

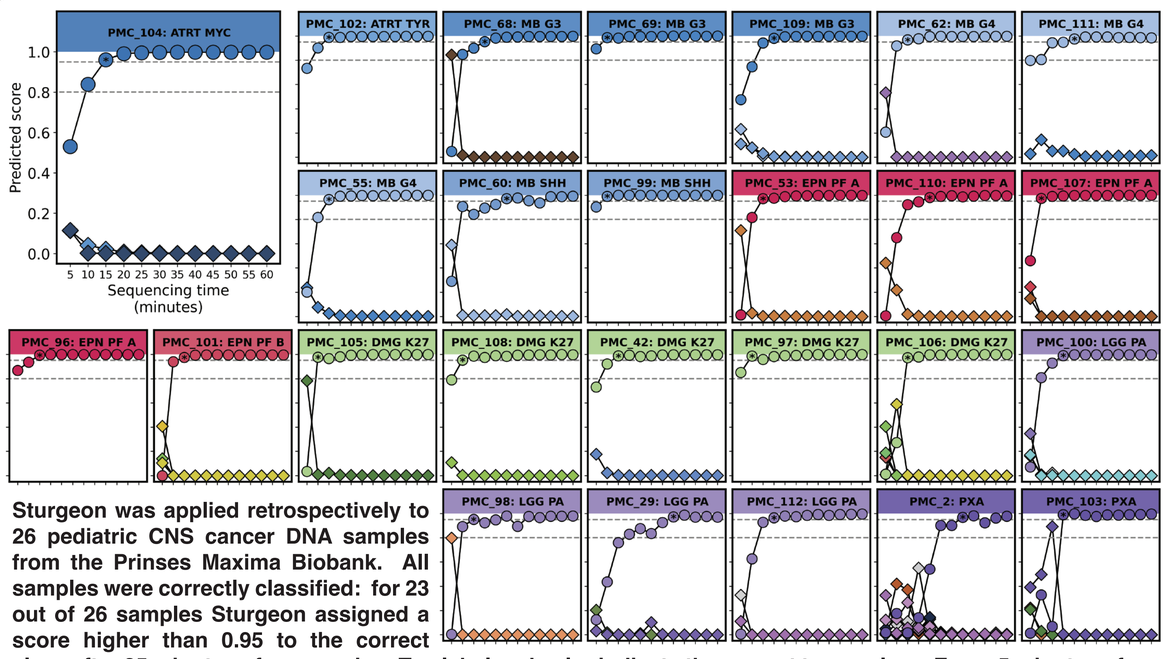
Our aim is to replace methylation arrays with shallow nanopore sequencing, which can be performed within 90 minutes, providing the molecular diagnosis in time to adjust the surgical strategy. The main challenge is to perform classification with only a few thousand random CpG sites, rather than the 450.000 sites obtained from a typical methylation array. To this end we developed Sturgeon, a machine learning framework for ultrafast CNS tumor diagnosis.

Sparsity problem



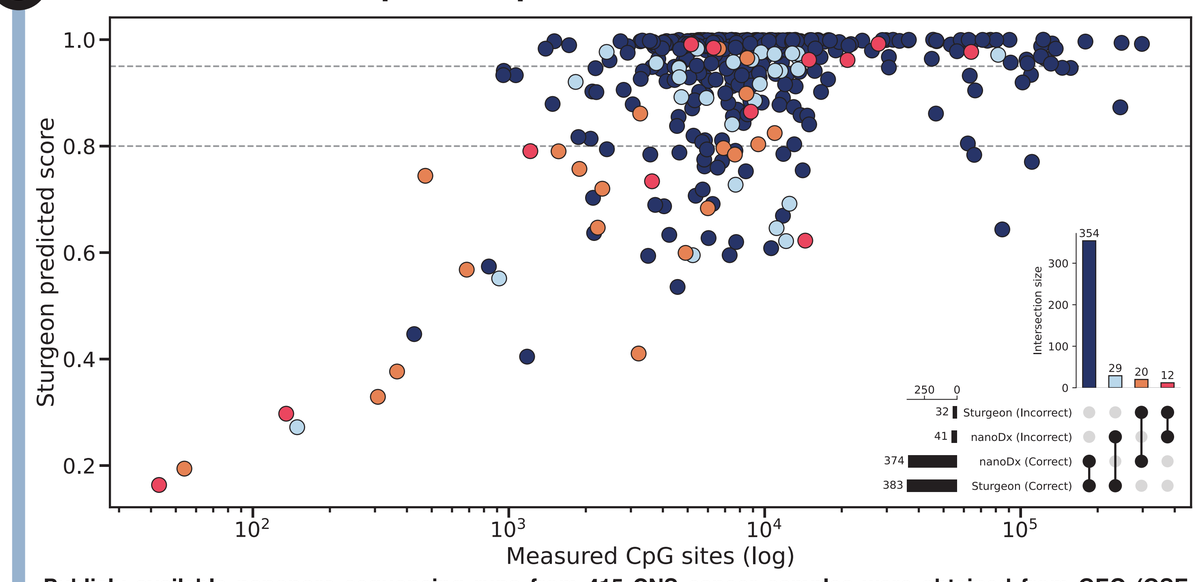
A full methylation array profile contains the methylation status of many CpG sites (~450k): here shown for a fraction of the sites (blue = unmethylated, orange = methylated, grey = unkonwn). Using nanopore sequencing, more and more reads are obtained over time, slowly filling up the profile. However, it is impossible to know *a priori*, which sites will be sequenced; and only a small fraction of the profile can be covered during surgery.

Validation on Nanopore sequenced pediatric CNS cancer samples



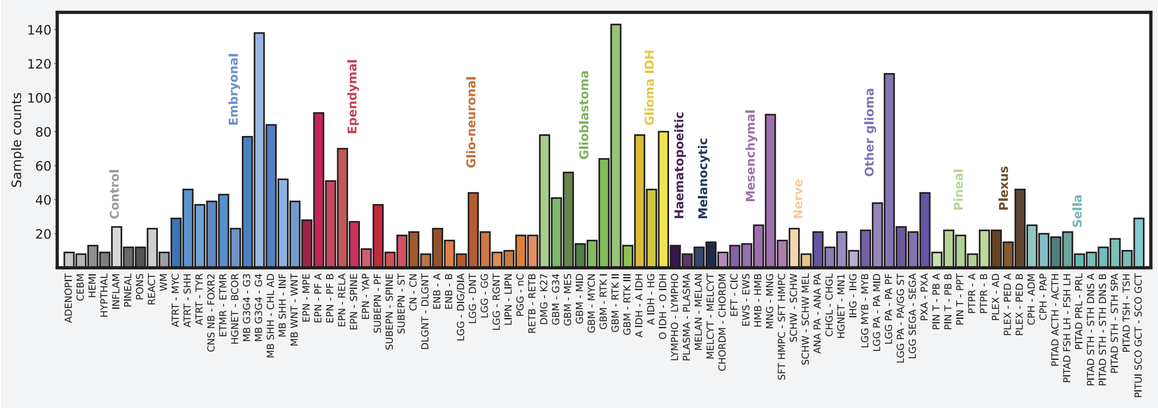
Sturgeon was applied retrospectively to 26 pediatric CNS cancer DNA samples from the Prinses Maxima Biobank. All samples were correctly classified: for 23 out of 26 samples Sturgeon assigned a score higher than 0.95 to the correct class after 25 minutes of sequencing. Top label and color indicate the correct tumor class. Every 5 minutes of sequencing a classification is made and scores are here represented: correct class (circle), incorrect class (diamond), asterisk indicates the first time the score was higher than 0.95, our classification threshold.

Validation on Nanopore sequenced external CNS cancer cohort



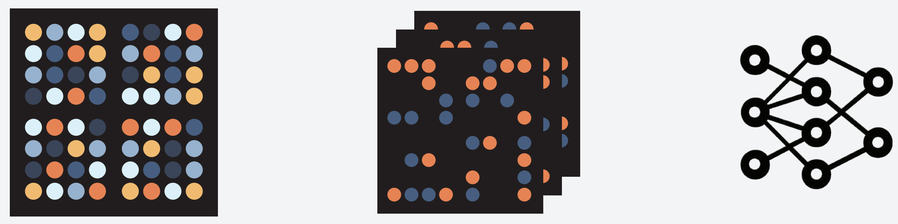
Publicly available nanopore sequencing runs from 415 CNS cancer samples were obtained from GEO (GSE 209865). We compare our model (Sturgeon), against nanoDx (Kuschel *et al.,* 2022), a patient-specific random forest classifier. Sturgeon correctly classified 383 (92.2%) samples, 343 (82.6%) at a threshold of 0.8 and 252 (60.7%) samples with a confidence >0.95. From the 415 samples, 32 (7.7%) were incorrectly classified of which 13 (3.1%) reached a confidence >0.8 and 8 (1.9%) reached a confidence score >0.95.

Reference data



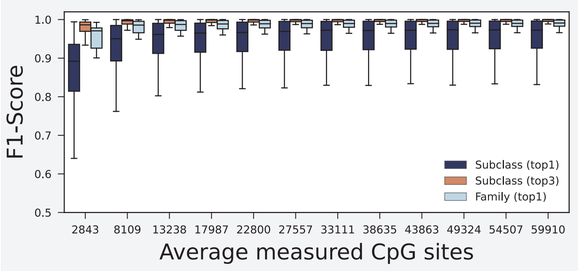
Sturgeon must first learn what CNS cancer methylation profiles look like. These are obtained from Illumina methylation arrays, available from GEO (GSE109381, Capper *et al.,* 2018) This dataset consists of methylation profiles from 2801 samples from 82 different cancer classes and 9 control tissues. Cancer classes are grouped into 14 larger family groups.

Simulation and model training

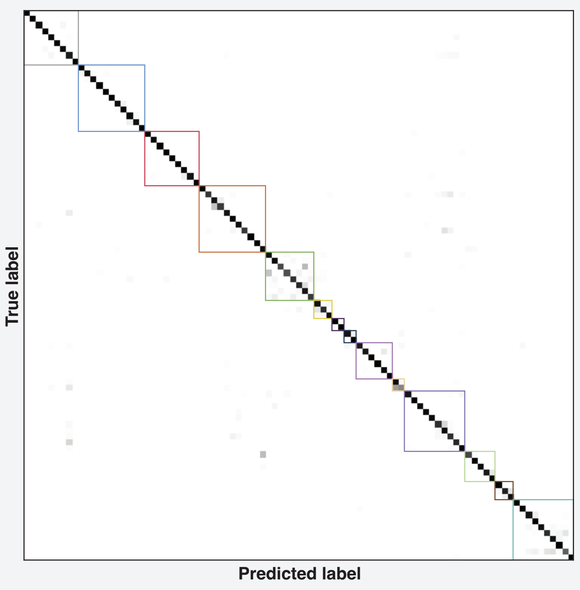


From the reference data, we simulate shallow nanopore sequencing runs (>36.000.000). In essense this means we randomly blind 85-99% of a methylation profile (while taking into account read parameters such as read length, sequencing depth, error rates, etc.). Since only a fraction of each profile is used, we can massively upsample from the 2801 full methylation profiles. These simulated nanopore runs are then used to train, calibrate and validate a neural network classifier.

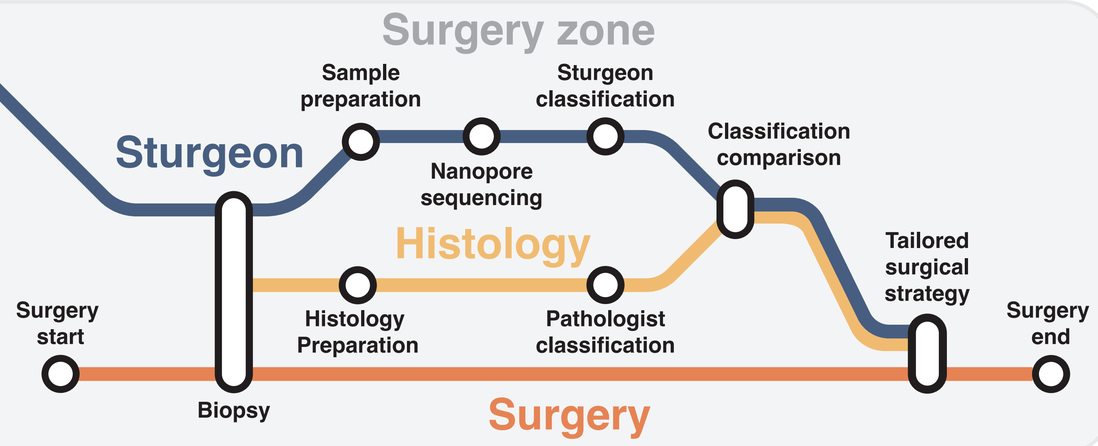
Validation on simulated data



We evaluate the performance of the neural network via 4-fold crossvalidation, here showing results for the test folds (25% of the reference dataset). Performance is sparsity dependant: median 0.85 F1-score with only 0.7% of the sites, 0.93 F1-score with 2.8% of the sites and 0.94 F1-score with 4.6% of the measured sites. These are the equivalent of 15, 35 and 55 minutes of sequencing. Most of the mistakes are within classes of the same family. The confusion matrix on the right shows model performance at 40 minutes of sequencing time (~97% missing data).



Methods zone



Conclusions

Sturgeon is a promising tool to perform intra-operative molecular classification of CNS tumors.

In retrospective testing, a correct diagnosis was reached in 45/49 cases with a maximum of 40 minutes of sequencing. 4 samples had inconclusive results.

We have performed 4 intraoperative analysis. Upon taking the biopsy, and taking sample preparation into account, we reached a diagnosis in less than 90 minutes.

The results from Sturgeon will be deployed in parallel to histological assessment and can be a valuable tool for pathologists performing histological assessment.

Sequencing can be continued, even after surgery, for additional data analysis such as copy number variation.

UMC Utrecht · Princess máxima center pediatric oncology · Oncode Institute

[1] Center for Molecular Medicine, University Medical Center Utrecht and Oncode institute
[2] Princess Máxima Center for Pediatric Oncology [3] Department of Pathology, Amsterdam University Medical Centers/VUmc
✉ j.deridder-4@umcutrecht.nl [†] These authors contributed equally