



**Vilnius
University**

Doktoranto Karolio Šablausko ataskaita už
2025/2026 mokslo metų antrąjį pusmetį

Darbo vadovė: prof. Audronė Jakaitienė

Disertacijos pavadinimas: Genetinių pokyčių charakterizavimas naudojant giliojo mokymo neuroninius tinklus (angl. *Characterization of genetic changes using deep neural networks*)

Darbo vadovas: prof. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2022 – 2027
(akademinės atostogos 2024-10-01 – 2025-01-31)

Studijų metai: 4.

Visų
doktorantūros
studijų ir
mokslinių tyrim
planas bei jo
vykdymo
suvestinė

1 lentelė: Doktorantūros studijų planas

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	1	1
II (2023/2024)	2	2
Akademinių atostogos 2024-10-01 – 2025-01-31		
III (2025/2026)	1	1
IV (2026/2027)	0	0
Iš viso:	4	4

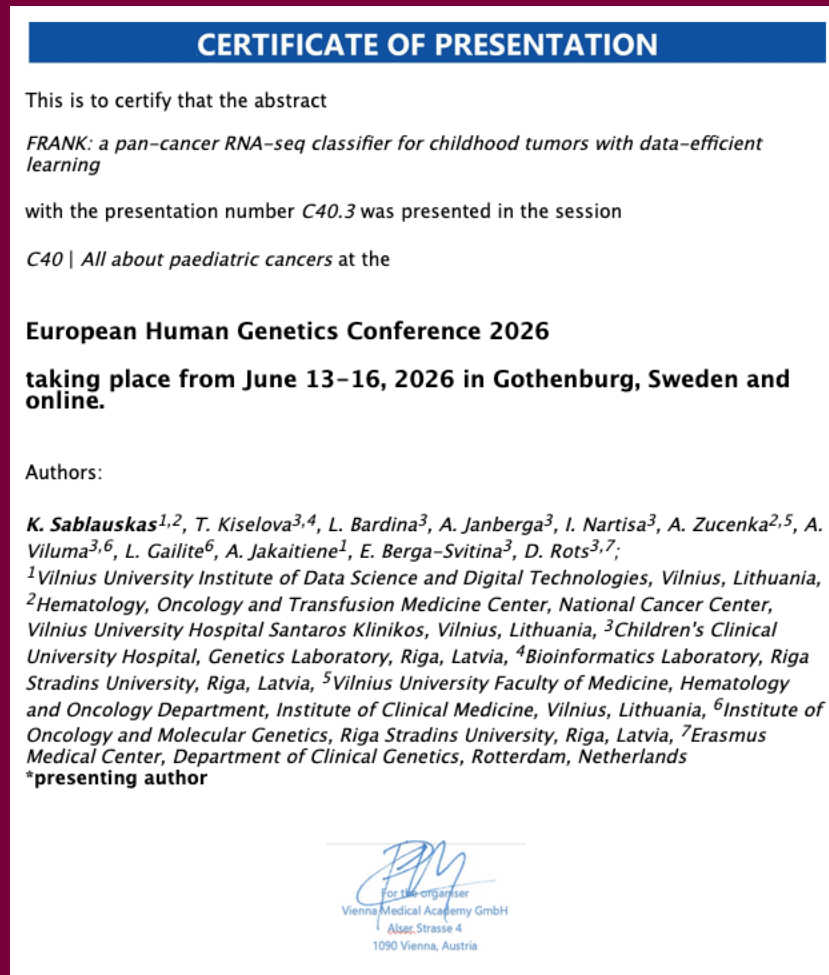
Visų doktorantūros studijų ir mokslinių tyrimų planas bei jo vykdymo suvestinė

Studijų metai	Dalyvavimas konferencijose				Publikacijos			
	Tarptautinėse		Nacionalinėse		Su citavimo rodikliu			Be citavimo rodiklio
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Publikacijų be citavimo rodiklio teikti nenumatoma
I (2022/2023)	0	0	0	0	0	0	0	
II (2023/2024)	1	0	0	0	0	0	0	
Akademinės atostogos 2024-10-01 – 2025-01-31								
III (2025/2026)	1	2 pristatymai	0	0	1	1	1 publikacija	
IV (2026/2027)	1	1 pristatymas	0	0	1	0	+ 1 rankraštis (pateikta Nature Communications – atmetė 2026-03-17 2026-03-18 pateikta Science Translational Medicine – atmetė 2026-06-30 Teikiama į žurnalą Bioinformatics	
Iš viso:	3	3	0	0	2	1	1	

Informacija apie tarptautinius renginius ir publikacijas (1)

Pristatytas žodinis pranešimas tarptautinėje konferencijoje „European Society of Human Genetics 2026“

Pranešimo tema: „FRANK: a pan-cancer RNA-seq classifier for childhood tumors with data-efficient learning“



Informacija apie tarptautinius renginius ir publikacijas (2)

Paruoštas rankraštis „FRANK: a pan-cancer RNA-seq classifier for childhood tumors with data-efficient learning” kartu su autoriais iš Latvijos.

2026-03-02 pateikta Nature Communications – atmetė
2026-03-17 be recenzijos.

2026-03-18 rankraštis pateiktas Science Translational
Medicine – atmetė 2026-06-30 po recenzijos.

2026-10 teikiama į Bioinformatics (IF 5,5). Patalpintas
rankraštis medrxiv



The screenshot shows the medRxiv preprint server interface. At the top is the medRxiv logo and the tagline 'THE PREPRINT SERVER FOR HEALTH SCIENCES'. Below this is a 'Follow this preprint' button. The title of the preprint is 'FRANK: a pan-cancer RNA-seq classifier for childhood tumors with data-efficient learning'. The authors listed are Karolis Sablauskas, Tatjana Kiselova, Livija Bardina, Inga Nartisa, Andrius Zucenka, Agnese Viluma, Linda Gailite, Audrone Jakaitiene, Egija Berga-Svitina, and Dmitrijs Rots. A DOI link is provided: <https://doi.org/10.64898/2026.09.23.26355605>. A disclaimer states: 'This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.' Navigation tabs include 'Abstract', 'Info/History', 'Metrics', and 'Preview PDF'. The 'Abstract' tab is selected, showing the following text: 'Accurate characterization of tumor subtype is of paramount importance in guiding treatment decisions in pediatric cancers. Here, we present FRANK (Fully-connected RNA-based Augmentation Neural Klassifier), a machine learning approach for classifying pan-cancer pediatric tumors. We curated a set of 11,467 transcriptomes to build a high-quality training dataset representing 181 distinct childhood tumor and normal tissue types. We used extensive data augmentation techniques with a hierarchical penalty term to train a neural network capable of detecting tumor types with as few as three training examples. To evaluate the generalization of the trained model, we perform validation on nine datasets spanning 16,398 samples, achieving an overall accuracy of 98.3%. Furthermore, we demonstrated that FRANK outperformed other published pediatric pan-cancer RNA-seq classifiers on an external dataset. To evaluate the robustness, we showed that FRANK remained accurate when applied to long-read RNA-sequencing data, as well as to noisy and low-quality data.'

Visų doktorantūros mokslinių tyrimų ir disertacijos rengimo etapai

3.	Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas: 1. Tyrimų apžvalga ir analizė. 2. Tyrimo metodikos sudarymas. 3. Teorinis tyrimas. 4. Empirinis tyrimas. 5. Gautų duomenų analizė, apibendrinimas. 6. Išvados, įvadas, literatūros sąrašas.	2026 balandžio mėn. – 2027 vasario mėn.	1. Publikacijos navikų tipo nustatymo tema teikimas 2. Apmokomi modeliai kopijų skaičiaus pokyčių aptikimui (daugiasluoksnis perceptronas, konvoliucinis neuroninis tinklas ir transformerio modeliai). 3. Atliekama modelių hiperparametrų paieška. 4. Publikacijos kopijų skaičiaus pokyčių aptikimui tema rengimas ir teikimas. 5. Disertacijos pateikimas svarstymui
----	---	---	---

1.1 Thesis aim (Tyrimo tikslas)

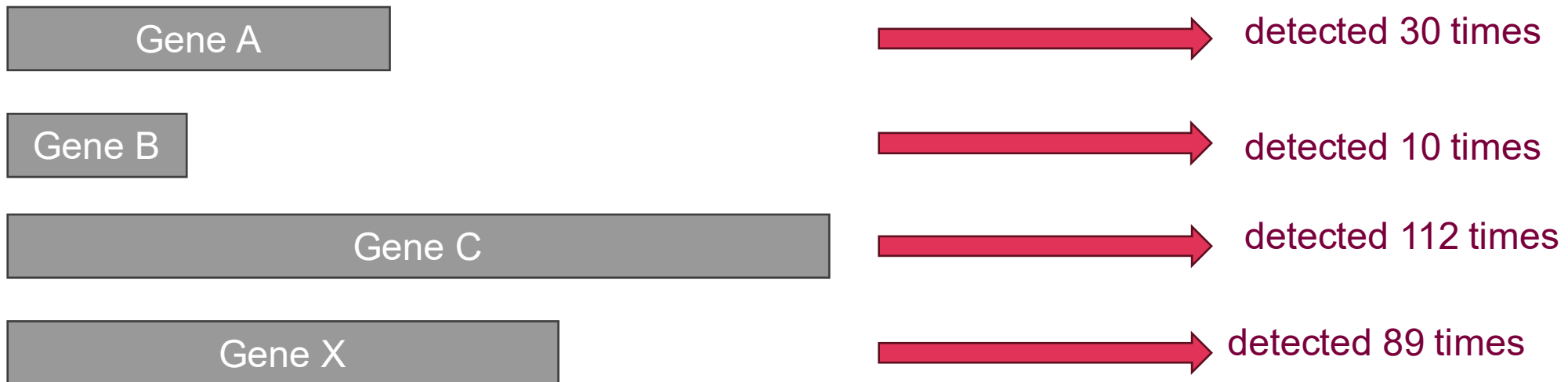
To contribute to the advancement of deep learning techniques for the analysis of next generation sequencing data, with a focus on RNA sequencing data.

1.2 Thesis objectives (Tyrimo uždaviniai)

- **Data preprocessing and feature engineering:** propose efficient data preprocessing pipeline suited for RNA sequencing data, including normalization and batch effect correction to prepare the data for further analysis.
- **Gene expression analysis using neural networks:** develop and implement neural network-based approaches for identifying gene expression patterns related to tumor types or their mutations.
- **Evaluation and benchmarking:** conduct extensive benchmarking and cross-validation experiments to assess the performance and generalizability of deep learning models, comparing them to traditional methods.
- **Biological case study:** apply said techniques in the interpretation of biological data gathered during a biomedical study.

Data object – whole transcriptome sequencing

The data object used in this work is RNA-sequencing vector of length n where n corresponds to the number of times that particular gene's RNA was detected. The values can integers (raw counts) or normalized.



Overview – tasks approached during this research

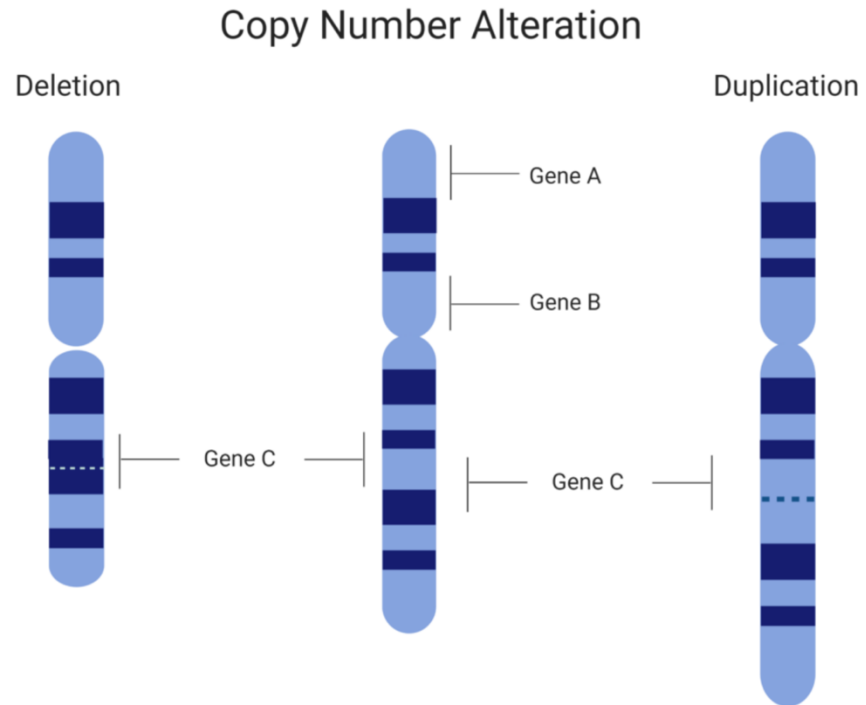
Copy-number aberration prediction using RNA-seq vector (GenoFormer model)

Final architecture: transformer

Tumor type prediction based on RNA-seq vector (FRANK model)

Final architecture: fully connected neural network

Copy number prediction task



<https://www.sourcekids.com.au/some-useful-terms-to-know-for-genetic-testing/>

Input: RNA-seq vector (length = n genes)

0.2	1.4	1.3	0.9	...	0.5	1.6	1.1	0.3
-----	-----	-----	-----	-----	-----	-----	-----	-----



Output: neutral (0), duplication or deletion for each chromosome arm (length = 79)

0	1	...	0	-1	0	0
---	---	-----	---	----	---	---

Each cohort (2 training/held-out split and 1 testing) was normalized independently:

1. log2-normalization
2. Per-gene subtraction of median of the cohort's normal karyotype (no aberrations) samples – training split only
3. Per-sample median subtraction
4. Clipping to $\pm 3\log 2$.

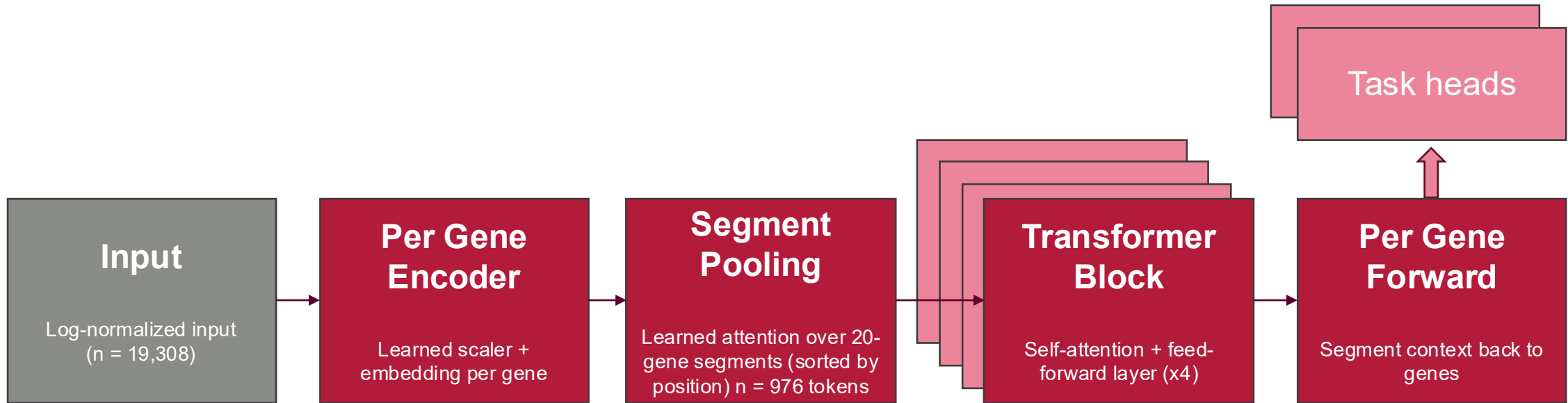
As ablation strategies during training:

1. Genes were binned into groups of 20 and a mean value was used.
2. Top- k genes by variance used as input
3. Average gene's expression across the cohort used as input

For copy number prediction task, the following strategies were tested:

1. Multilayer perceptron (MLP) with segment or full gene list as input
2. 1D-convolutional neural networks (CNNs) with segments or full gene list as input and Optuna hyperparameter search
3. Transformer with segment or full gene list as input and masked-pretraining

Transformer architecture



Masked pretraining

Unlabeled sample

0.2	1.4	1.3	0.9	0.5	0.5	1.6	1.1	0.3
-----	-----	-----	-----	-----	-----	-----	-----	-----

Unlabeled masked sample

0.2	1.4	x	0.9	0.5	0.5	x	x	0.3
-----	-----	---	-----	-----	-----	---	---	-----



Transformer



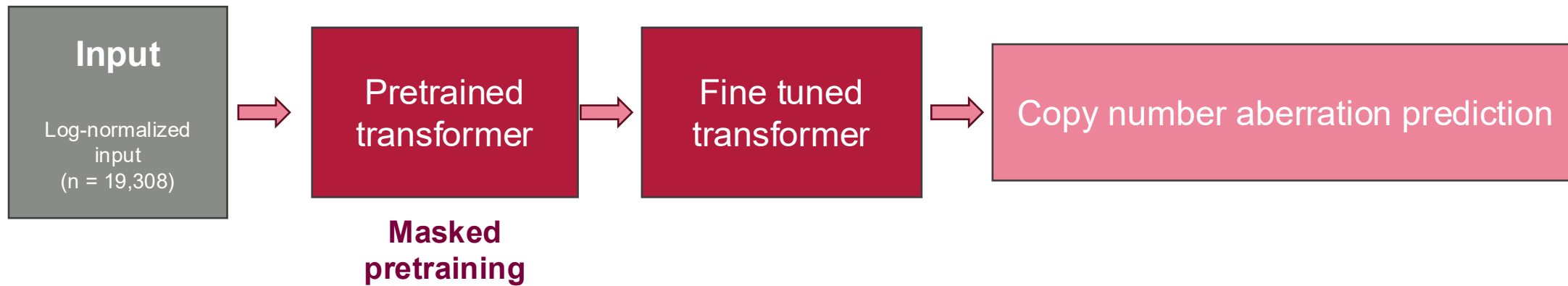
Reconstructed sample

0.2	1.4	1.2	0.9	0.5	0.5	1.4	1.0	0.3
-----	-----	-----	-----	-----	-----	-----	-----	-----

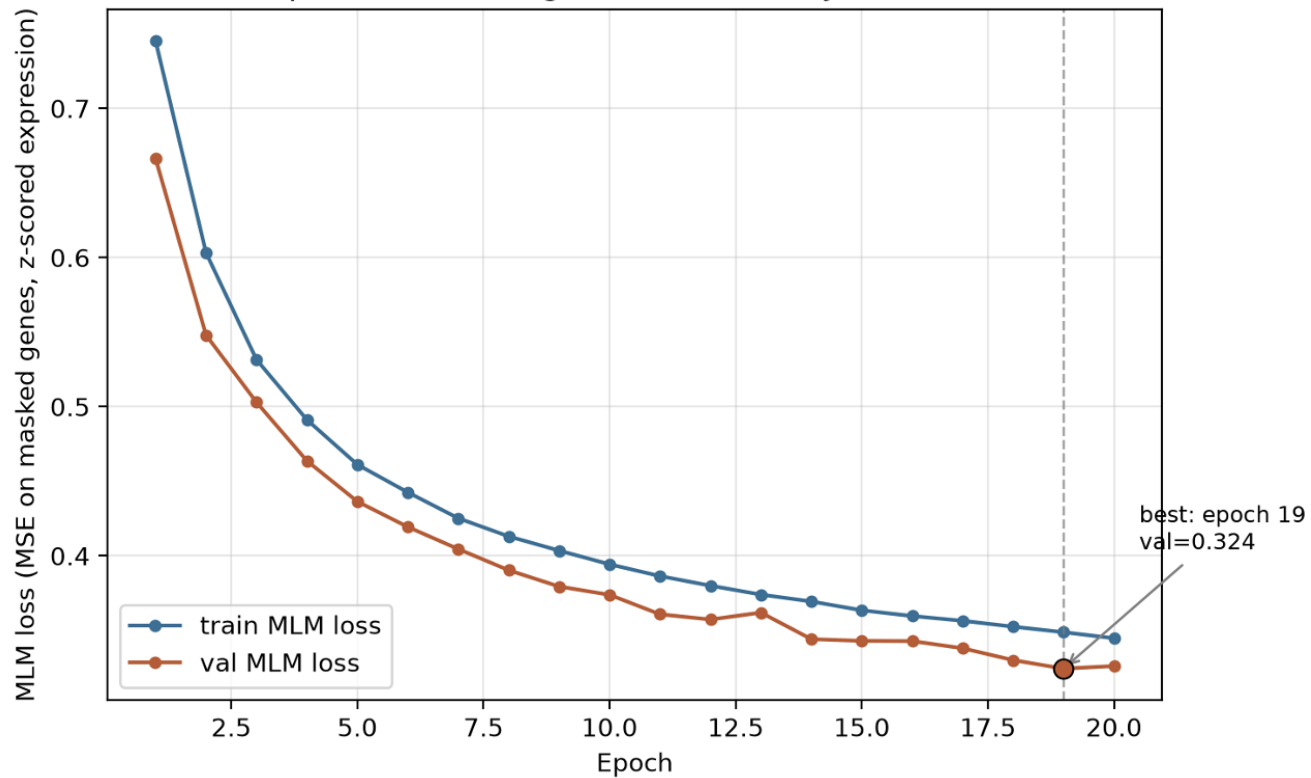
Masked pretraining uses a pool of unlabelled data to train reconstruction of masked values.

Loss (MSE) calculated on masked positions only.

Training workflow



Pretraining results



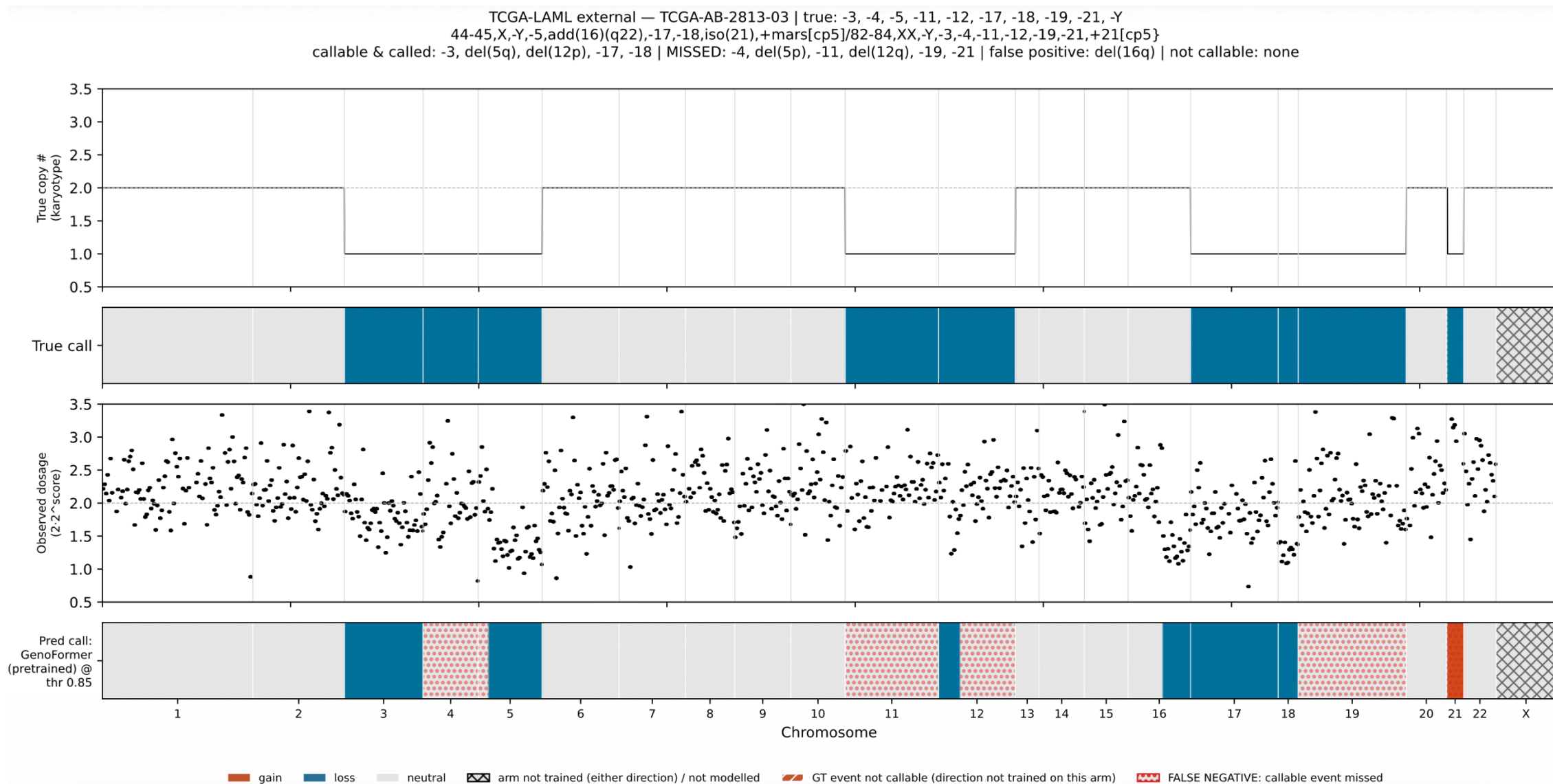
Pretraining metric: masked-expression-modeling (MLM) – masked MSE, 15% of each sample's genes are randomly masked, and the model predicts their original.

Number of model parameters 3.57M.

Comparison of architectures

Family	Model	Heldout AUROC	External AUROC	F1 @0.5	Calibrated Threshold	Cal. F1
MLP	gene-level MLP (all genes)	0,898125	0,948045	0,40146	0,55	0,417755
1D-CNN	gene-level 1D-CNN, variant 'all' (per-arm scale + hybrid arm statistics + per-gene feature channels)	0,88577	0,933655	0,343783	0,91	0,566406
Transformer	GenoFormer, pretrained backbone + per-arm head	0,930687	0,961341	0,436363	0,85	0,610837
Ensemble	GenoFormer (pretrained) + segment transformer d128 L4, mean of held-out-standardised logits	0,935516	0,966516		0,95	0,622391

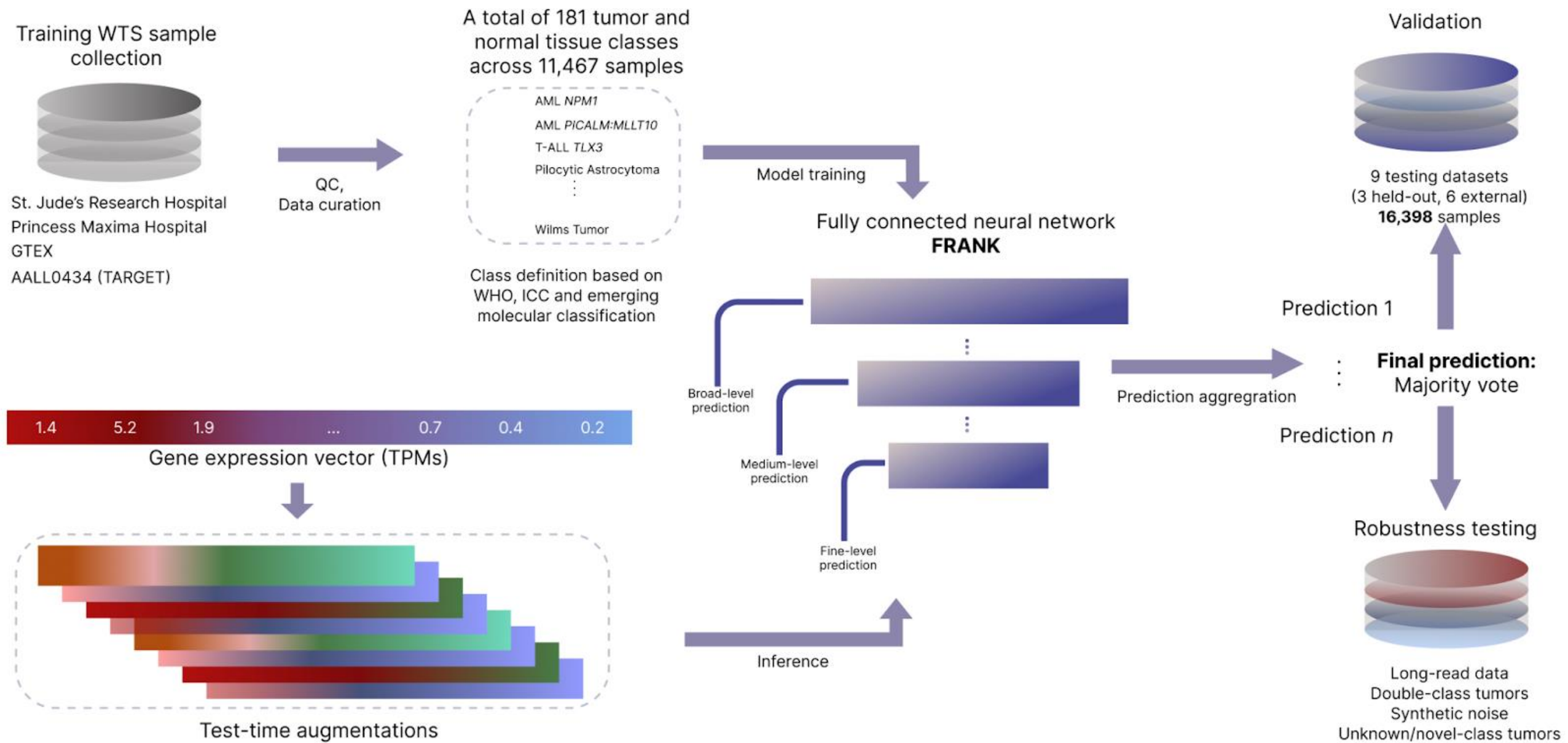
Example of model prediction





Part 2

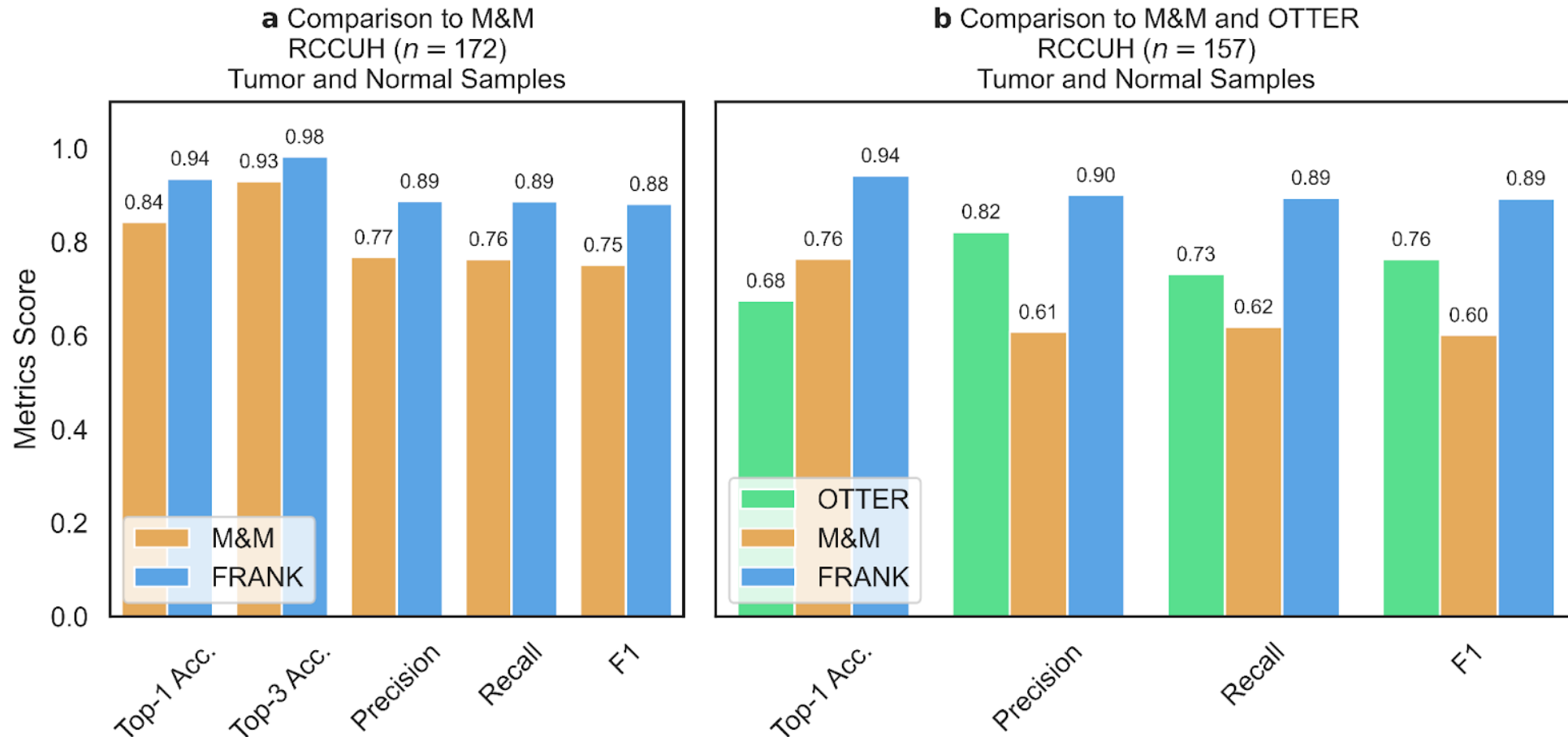
Predicting tumor type from RNA-seq data



Results in testing cohorts

Dataset	External dataset	N samples	N errors	N tumor/normal mismatches	Accuracy (%)	F1 score	Precision	Recall
RCCUH	Yes	178	13	1	92.7	0.87	0.88	0.88
Princess Maxima	No	430	21	4	95.1	0.90	0.92	0.90
TARGET AML	Yes	489	42	7	91.4	0.80	0.87	0.77
TARGET Other	Yes	468	32	1	93.2	0.90	0.99	0.84
AALL0434	No	707	37	0	94.8	0.85	0.90	0.82
Saint-Louis T-ALL	Yes	72	7	0	90.3	0.87	0.86	0.91
Beat AML	Yes	329	49	5	85.1	0.72	0.72	0.79

Comparison with other models



Kito semestro planas

- Publikuotas navikų klasifikacijos (FRANK) rankraštis (Bioinformatics)
- Pateiktas publikacijai antras rankraštis (GenoFormer)
- Pateiktas disertacijos rankraštis