Vilnius University

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

Darbo vadovė: prof. Audronė Jakaitienė

Vilnius University

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
Akademinės atostogos 2024-10-01 – 2025-01-31		
III (2025/2026)	**1**	**1**
IV (2026/2027)	0	0
Iš viso:	**4**	**4**

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

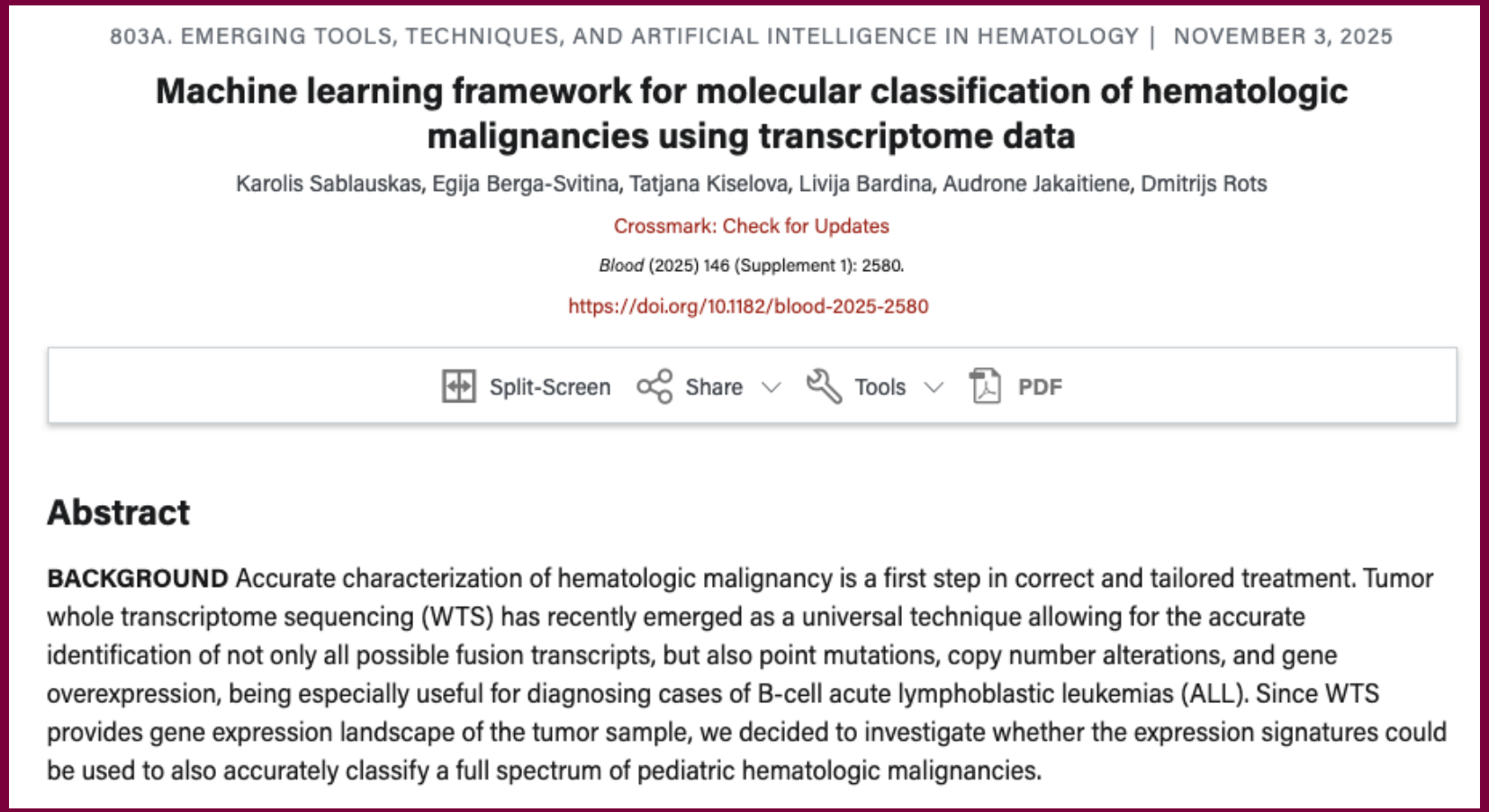
Vilnius University

Informacija apie tarptautinius renginius ir publikacijas (1)

Vilnius University

Pristatytas stendinis pranešimas tarptautinėje konferencijoje „American Society of Hematology Annual Meeting 2025“

Pranešimo tema: „ Machine learning framework for molecular classification of hematologic malignancies using transcriptome data“



803A. EMERGING TOOLS, TECHNIQUES, AND ARTIFICIAL INTELLIGENCE IN HEMATOLOGY | NOVEMBER 3, 2025

Machine learning framework for molecular classification of hematologic malignancies using transcriptome data

Karolis Sablauskas, Egija Berga-Svitina, Tatjana Kiselova, Livija Bardina, Audrone Jakaitiene, Dmitrijs Rots

Crossmark: Check for Updates

Blood (2025) 146 (Supplement 1): 2580.

https://doi.org/10.1182/blood-2025-2580

Split-Screen Share Tools PDF

Abstract

BACKGROUND Accurate characterization of hematologic malignancy is a first step in correct and tailored treatment. Tumor whole transcriptome sequencing (WTS) has recently emerged as a universal technique allowing for the accurate identification of not only all possible fusion transcripts, but also point mutations, copy number alterations, and gene overexpression, being especially useful for diagnosing cases of B-cell acute lymphoblastic leukemias (ALL). Since WTS provides gene expression landscape of the tumor sample, we decided to investigate whether the expression signatures could be used to also accurately classify a full spectrum of pediatric hematologic malignancies.

Informacija apie tarptautinius renginius ir publikacijas (2)

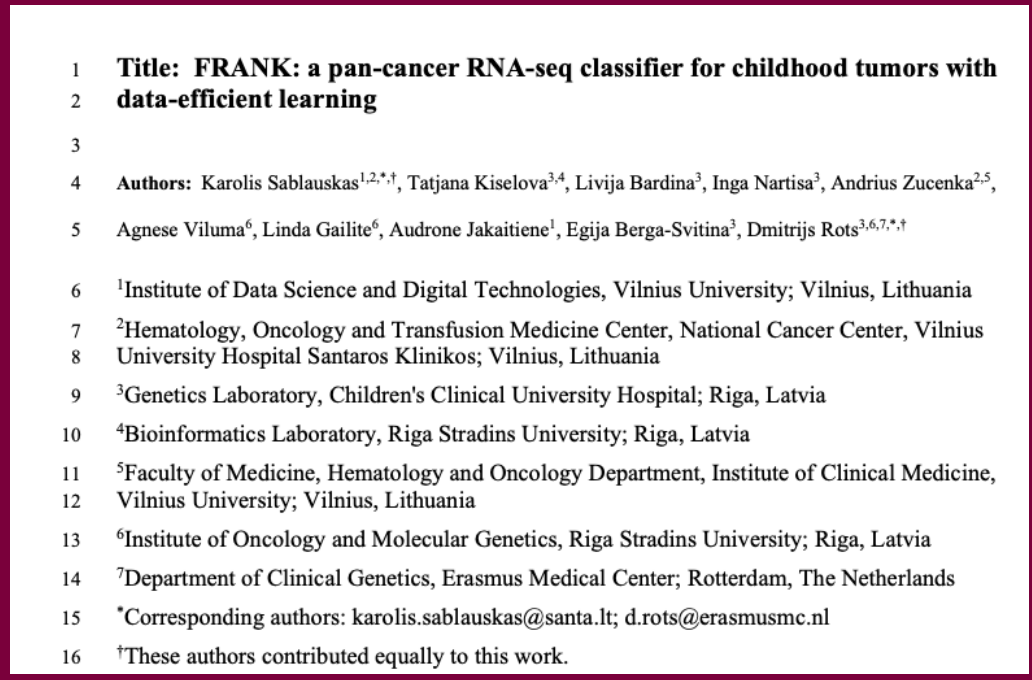
Vilnius University

Žurnale „Science“ (IF 49,7) publikuota „*High-throughput single cell -omics using semi-permeable capsules*“ kartu su autoriais iš VU Gyvybės mokslo centro.

Informacija apie tarptautinius renginius ir publikacijas (3)

Vilnius University

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



Title: FRANK: a pan-cancer RNA-seq classifier for childhood tumors with data-efficient learning

Authors: Karolis Sablauskas[1,2,*,†], Tatjana Kiselova[3,4], Livija Bardina[3], Inga Nartisa[3], Andrius Zucenka[2,5], Agnese Viluma[6], Linda Gailite[6], Audrone Jakaitiene[1], Egija Berga-Svitina[3], Dmitrijs Rots[3,6,7,*,†]

[1]Institute of Data Science and Digital Technologies, Vilnius University; Vilnius, Lithuania

[2]Hematology, Oncology and Transfusion Medicine Center, National Cancer Center, Vilnius University Hospital Santaros Klinikos; Vilnius, Lithuania

[3]Genetics Laboratory, Children's Clinical University Hospital; Riga, Latvia

[4]Bioinformatics Laboratory, Riga Stradins University; Riga, Latvia

[5]Faculty of Medicine, Hematology and Oncology Department, Institute of Clinical Medicine, Vilnius University; Vilnius, Lithuania

[6]Institute of Oncology and Molecular Genetics, Riga Stradins University; Riga, Latvia

[7]Department of Clinical Genetics, Erasmus Medical Center; Rotterdam, The Netherlands

[*]Corresponding authors: karolis.sablauskas@santa.lt; d.rots@erasmusmc.nl

[†]These authors contributed equally to this work.

2026-03-02 Pateikta Nature Communications – atmetė 2026-03-17. Redaktorius nurodė, jog galima perkelti į Communications Medicine (Nature grupės žurnalas, IF 6.3) ir straipsnis bus siunčiamas recenzijai.

2026-03-18 rankraštis pateiktas Science Translational Medicine (IF 14.6), laukiama atsakymo. Jeigu bus gautas neigiamas atsakymas, bus sprendžiama dėl perkėlimo Communications Medicine.

Informacija apie tarptautinius renginius ir publikacijas (3)

Vilnius University

Paruošto rankraščio duomenimis pateikta santrauka „European Society of Human Genetics 2026“ konferencijai.

Konferencijos potemė: “*Bioinformatics, statistical methods and* AI”

Control/Tracking Number: 2026-A-2583-ESHG
Activity: ESHG Abstract
Current Date/Time: 3/20/2026 3:54:45 AM

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

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.	2025 gruodžio mėn. – 2026 rugsėjo mėn.	Atliekama modelio hiperparametrų paieška taikant optuna, atliekama modelio validacija naudojant išsamius išorinius duomenis (viso 16 398 testavimo mėginiai), atliekama modelio “tvirtumo” (angl. robustness) analizė, testavimas naudojant klinikinius duomenis. Paruošiamas ir publikacijai pateikiamas rankraštis. Remiantis rankraščio duomenimis pateikiama santrauka konferencijai.

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)

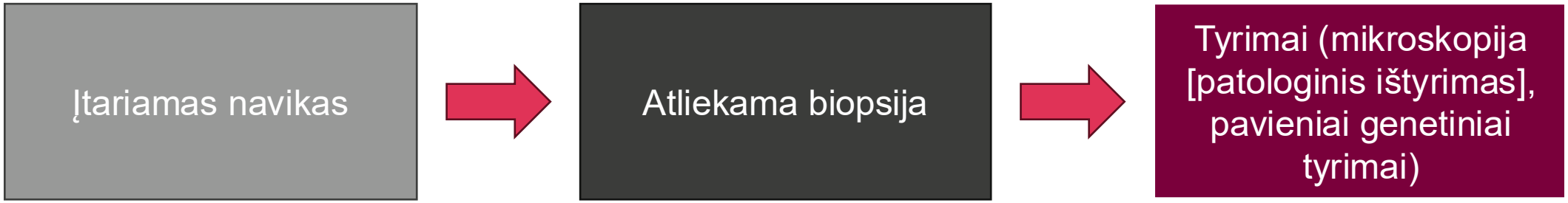
- **Data preprocessing and feature engineering:** create efficient data preprocessing pipeline suited for RNA sequencing data, including normalization and batch effect correction to prepare the data for further analysis.

- **Deep differential expression analysis:** develop and implement deep learning-based approach for identifying differentially expressed genes and pathways.

- **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.

Mokslinių rezultatų pristatymas: tyrimo problematika



Tikslas: konkrečiai įvardinti naviko tipą

Trukmė: 1-4 sav.

Vilnius University

Tyrimo problematika

Pavienių diagnostinių tyrimų trūkumai:

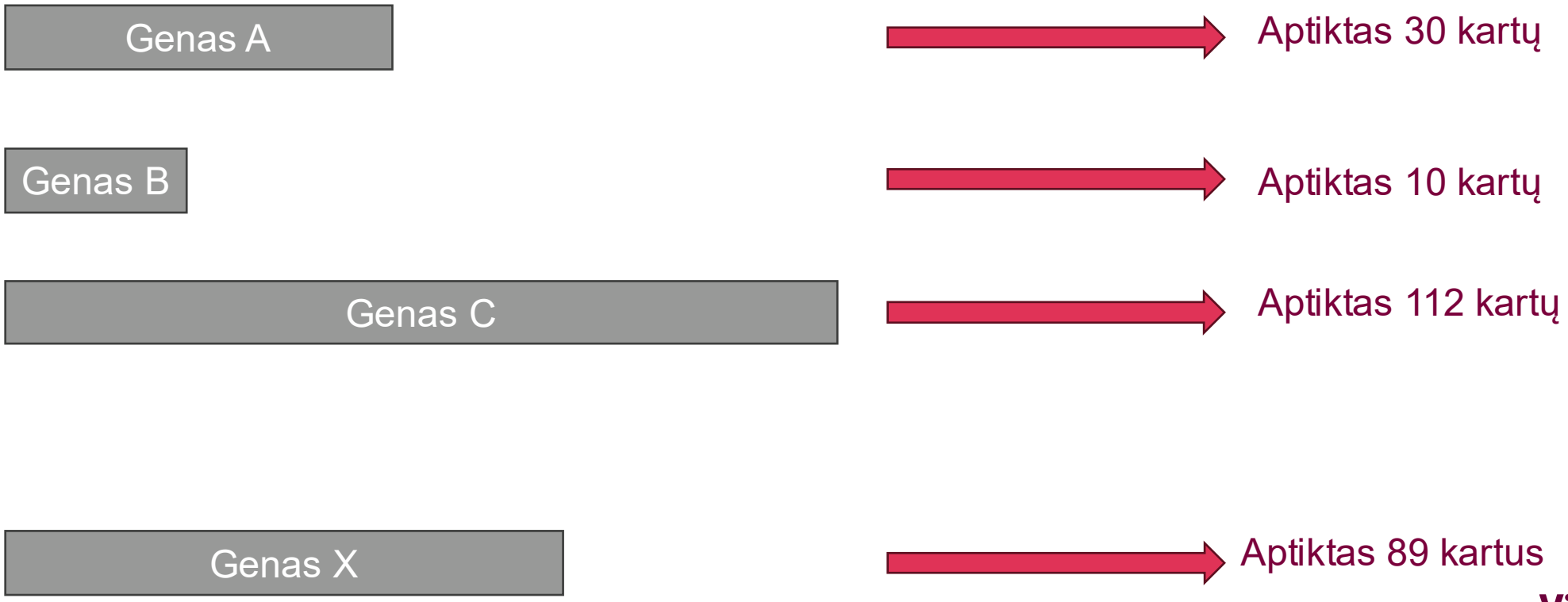
- **Sudėtinga iš infrastruktūros pusės (kiekvienam navikui reikia atskiro lab. metodo)**
- **Laiko trukmė gauti rezultatą.**

Galimas sprendimas:

- “omikos” metodai – didelio kiekio genų ištyrimas

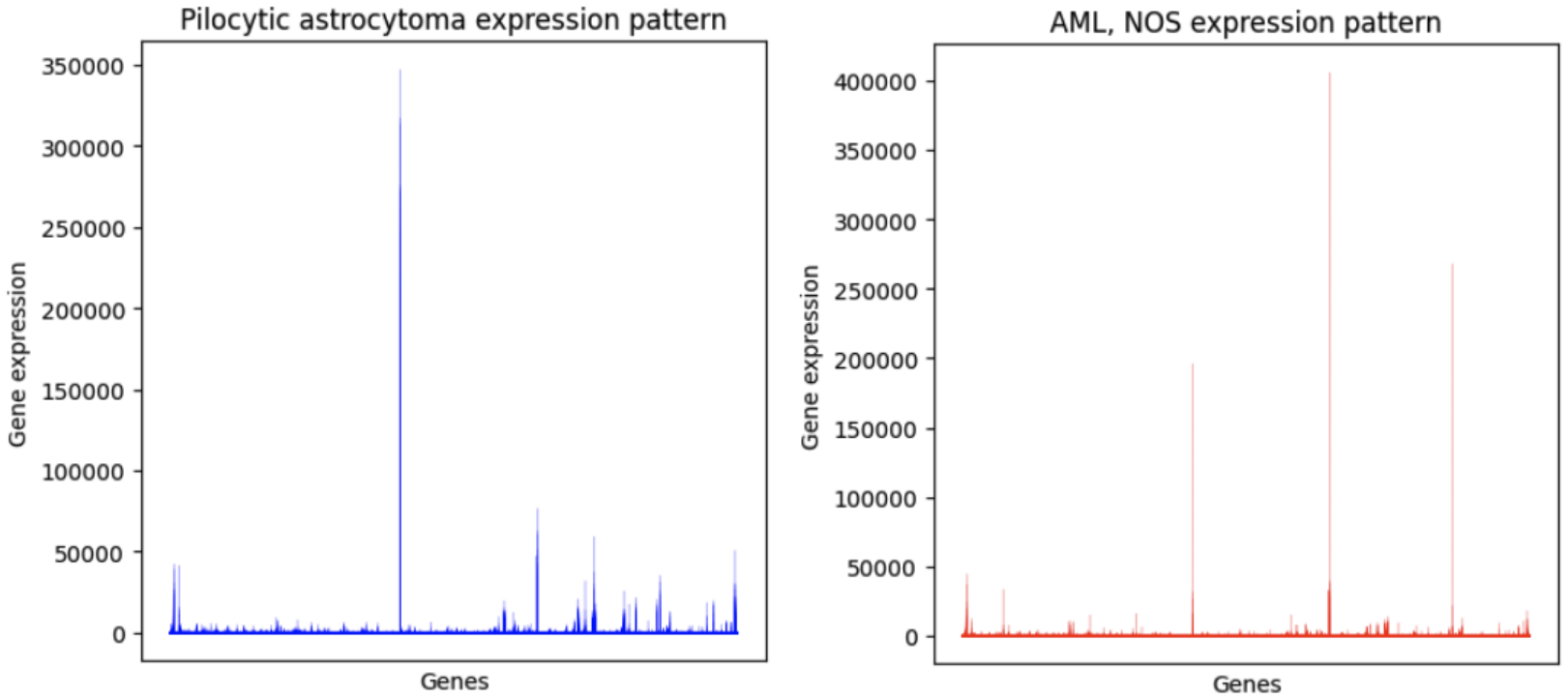
Vilnius University

Viso transkriptomo sekoskaita (angl. Whole transcriptome sequencing - WTS)



Vilnius University

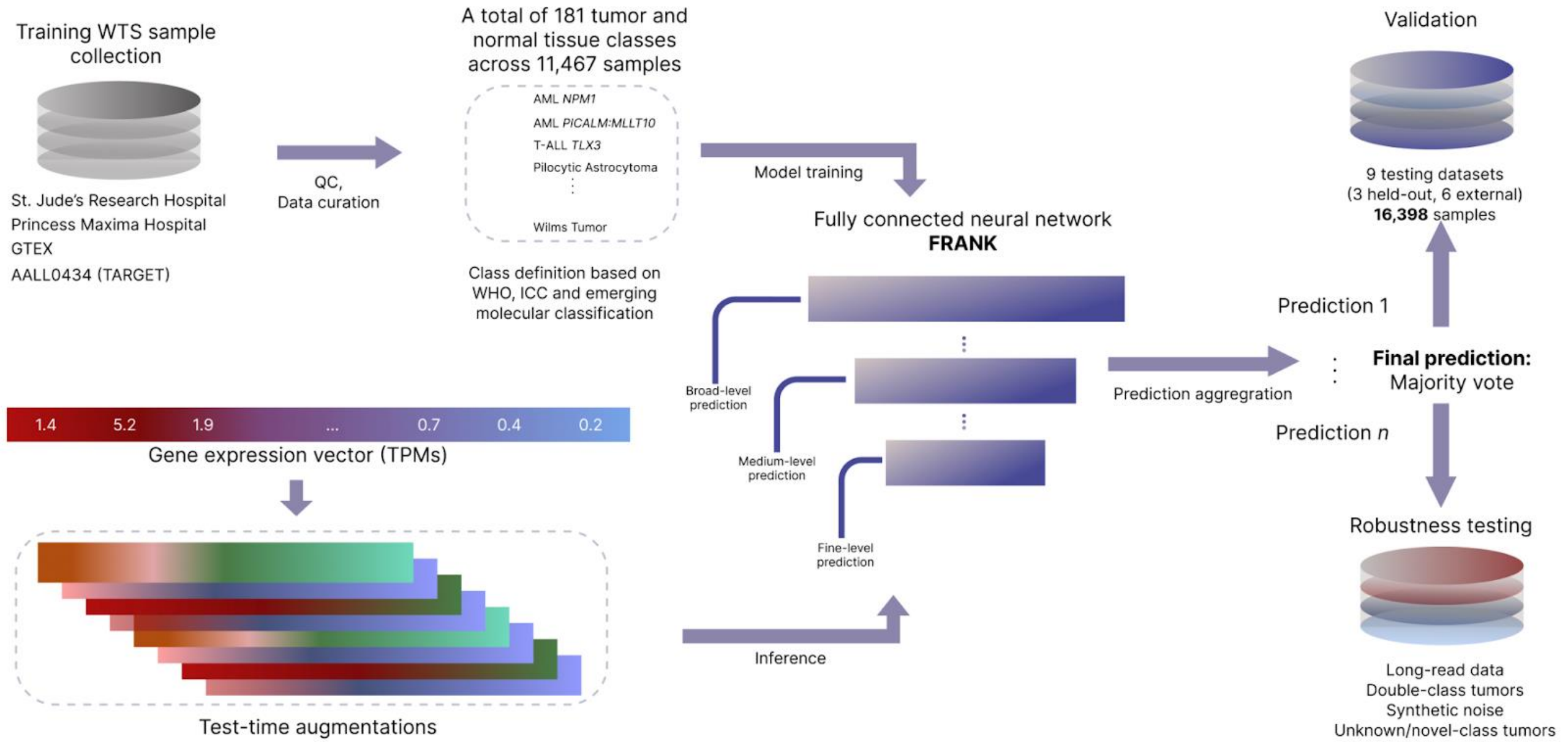
Genų raiškos palyginimas



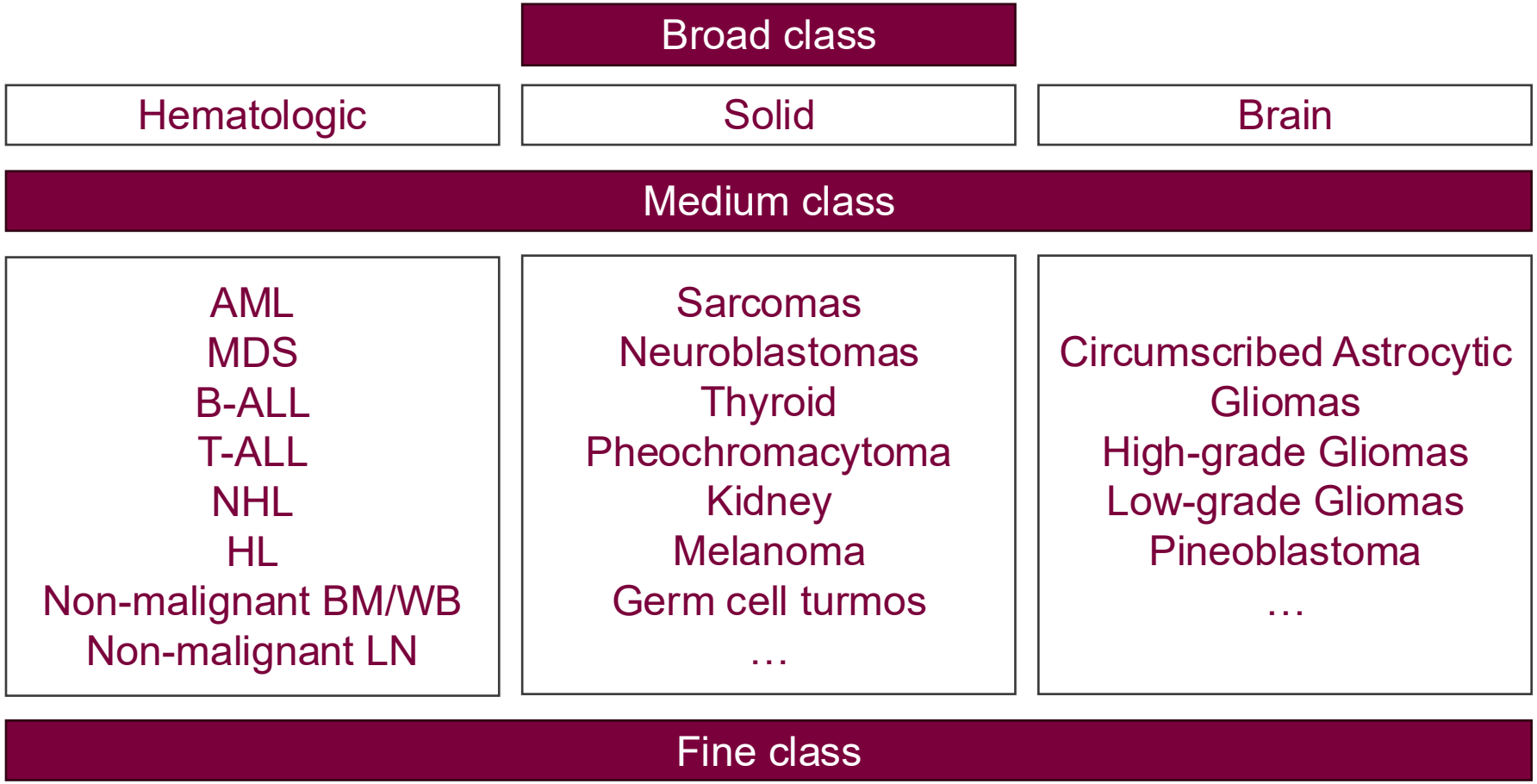

```
[ 0.         , 20.16973919,  0.         , ...,  0.3646838 ,
```

Using WTS gene expression to predict a specific malignancy subtype

No direct detection of fusion, mutations, etc.


Training WTS sample collection
St. Jude's Research Hospital
Princess Maxima Hospital
GTEX
AALL0434 (TARGET)
QC,
Data curation
A total of 181 tumor and normal tissue classes across 11,467 samples
AML NPM1
AML PICALM:MLLT10
T-ALL TLX3
Pilocytic Astrocytoma
Wilms Tumor
Class definition based on WHO, ICC and emerging molecular classification
Model training
Fully connected neural network
FRANK
Broad-level prediction
Medium-level prediction
Fine-level prediction
Prediction aggregration
Prediction 1
Prediction n
Final prediction:
Majority vote
Validation
9 testing datasets
(3 held-out, 6 external)
16,398 samples
Robustness testing
Long-read data
Double-class tumors
Synthetic noise
Unknown/novel-class tumors
1.4
5.2
1.9
...
0.7
0.4
0.2
Gene expression vector (TPMs)
Test-time augmentations
Inference

Hierarchinė navikų klasifikacija



Viso 181 klasė (navikai ir sveikas audinys)

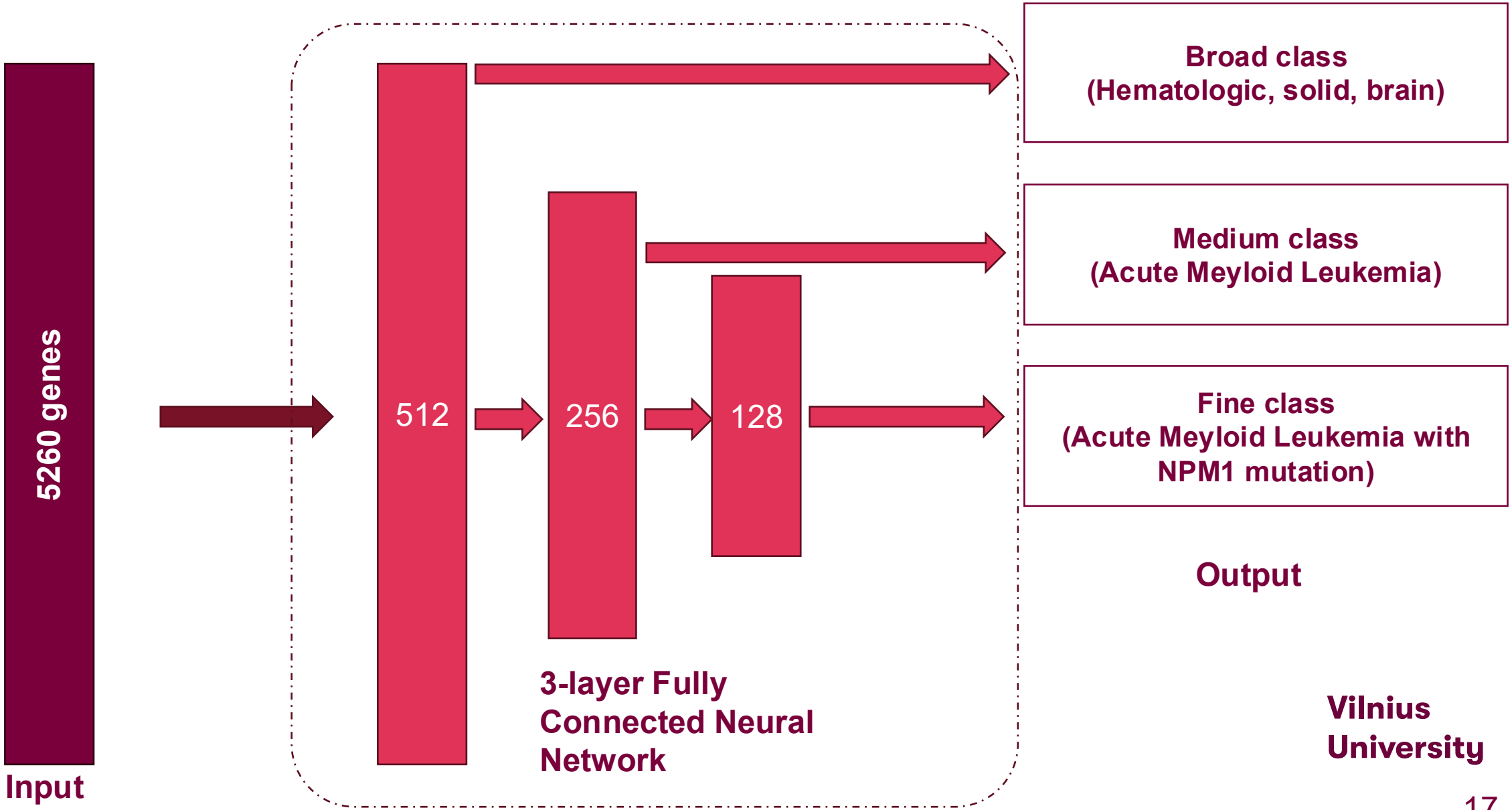
Vilnius University

Mokymo duomenų apžvalga

Vilnius University

Cohort	N individuals	N samples	N classes (fine level)	Tissue Preservative	Enrichment Method	GENCODE Version
St. Jude	3,911	4,319 (Tumor *n*=4319) (Normal n=0)	144 (Pan-tumor)	FFPE (*n*=805) Fresh/Stabilized (*n*=2,973) Not Available (*n*=541)	Not Available (*n*=5) PolyA (*n*=889) Total RNA (*n*=3,425)	v31
Princess Maxima	1,136	1,149 (Tumor *n*=1110) (Normal *n*=39)	80 (Pan-tumor)	Fresh/Stabilized (*n*=1149)	Total RNA (*n*=1,149)	v29
GTEx v10 release	283	5,714 (Tumor *n*=0) (Normal *n*=5,714)	31 (All normal tissue)	Fresh/Stabilized (*n*=5714)	Total RNA (*n*=5,714)	v39
AALL0434	285	285 (Tumor *n*=285) (Normal *n*=0)	7 (T-ALL)	Fresh/Stabilized (*n*=285)	Total RNA (*n*=285)	N.A.
Total	5,615	11,467 (Tumor *n*=5,714) (Normal *n*=5,753)	181			

Klasifikatorius



Vilnius University

Original image

Flip

Rotation

Blur

Exposure

Contrast

Grayscale

Augmented images

Duomenų papildymas

Triukšmas viso mėginio lygiu

120	**455**	**…**	**13**
119	457	…	15
Gene A	**Gene B**	**…**	**Gene X**

Trūkstami genai

120	**455**	**…**	**13**
120	0	…	13
Gene A	**Gene B**	**…**	**Gene X**

Triukšmas vieno geno lygiu

120	**455**	**…**	**13**
120	**677**	…	13
Gene A	**Gene B**	**…**	**Gene X**

CutMix

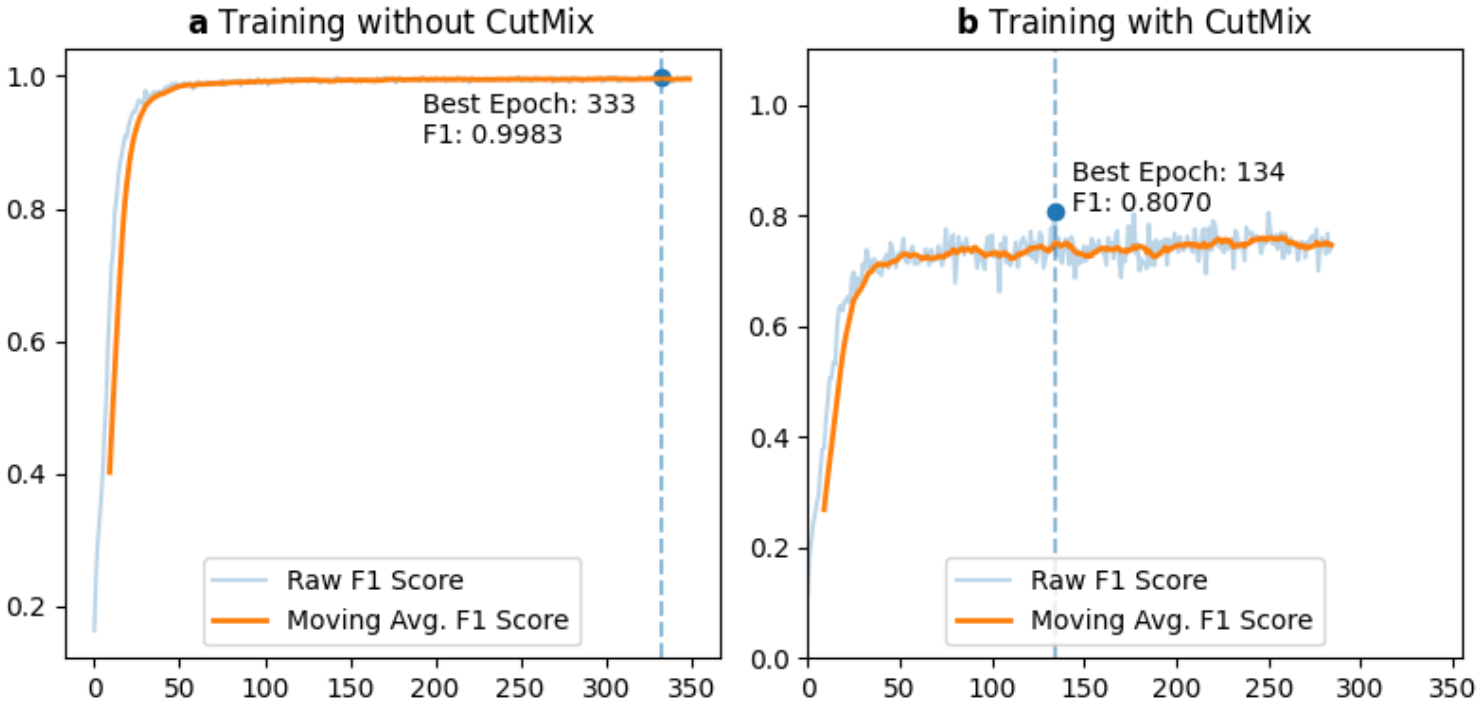
120	**455**	**…**	**13**
256	**567**	**…**	**8**

120	455	**…**	**8**

Vilnius University

Mokymo pasirinkimas

- Daugiaklasės klasifikacijos uždavinys - 181 klasė
- Klasių disbalansas (3 – 303 mokymų pavyzdžių klasei)
- Mokymas naudojant visus duomenis – *CutMix* padeda išvengti persimokymo.



Vilnius University

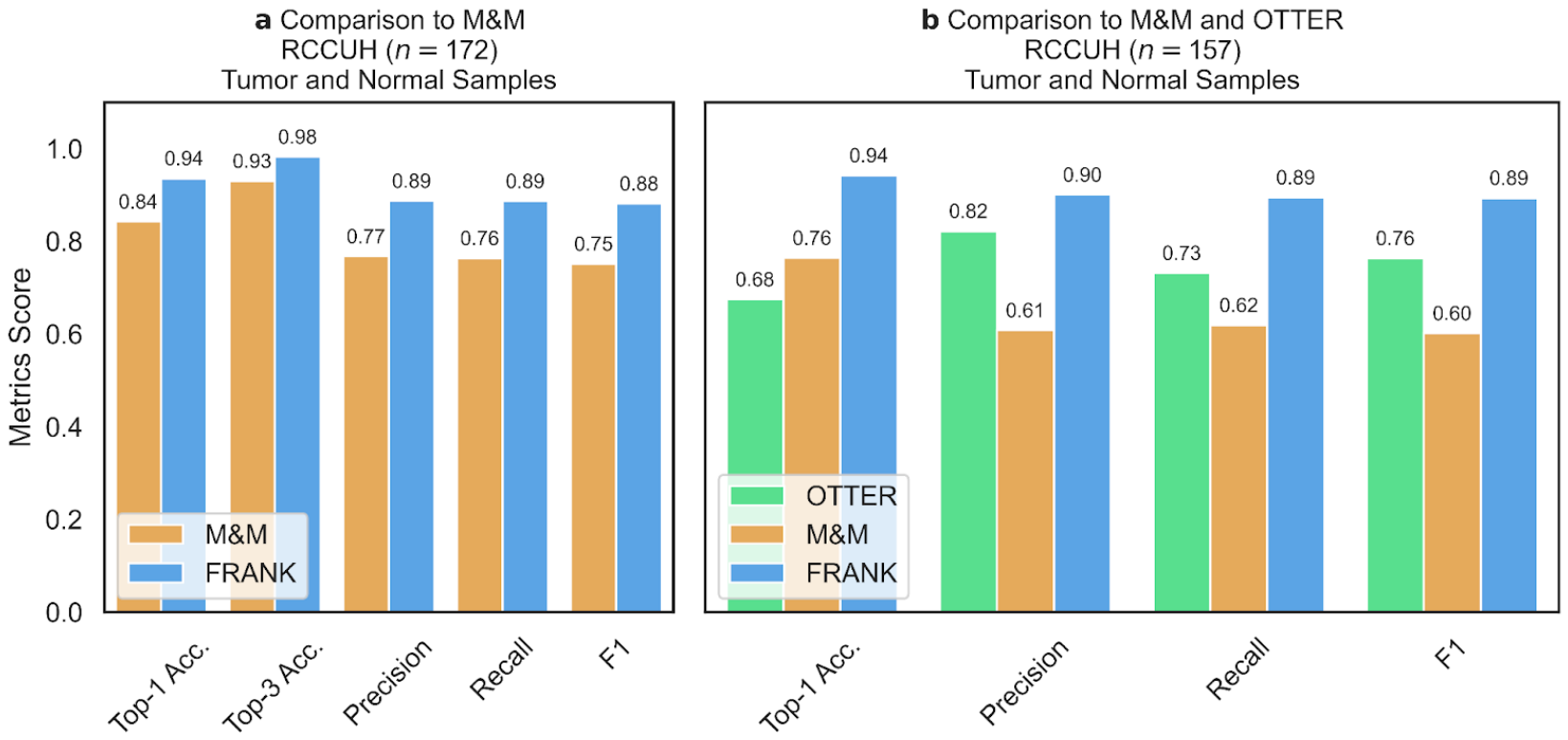
Rezultatai testavimo kohortose

Vilnius University

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

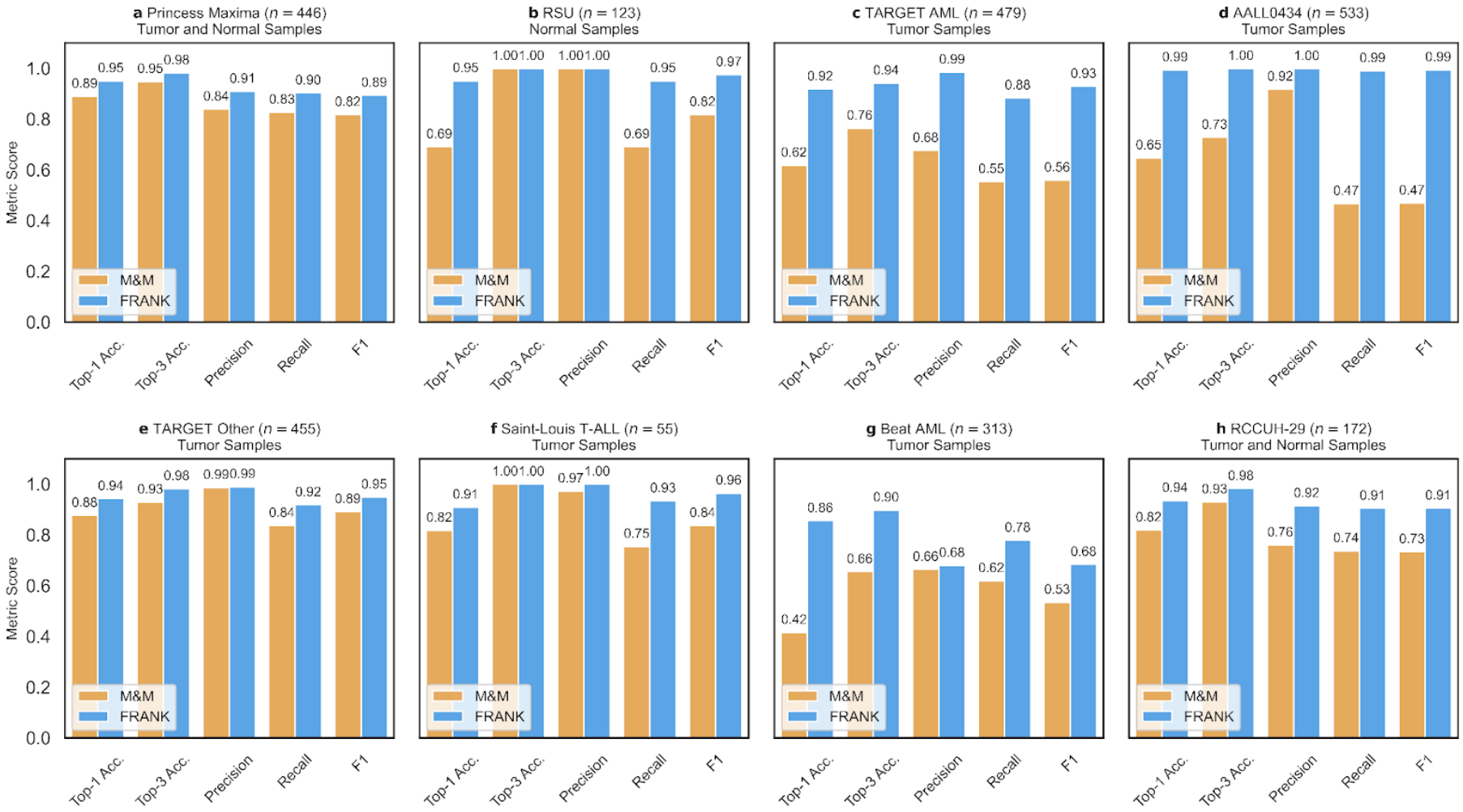
Palyginimas su kitais literatūroje aprašytais modeliais

- M&M ir OTTER
- Rygos ligoninės (RCCUH) duomenų rinkinys (nenaudotas publikacijose, patikimi ir “išvalyti” duomenys)
- M&M 172/189 mėginiai tinkami testavimui, OTTER 157/189 mėginiai tinkami testavimui (visi tinkami OTTER, tinkami ir M&M)



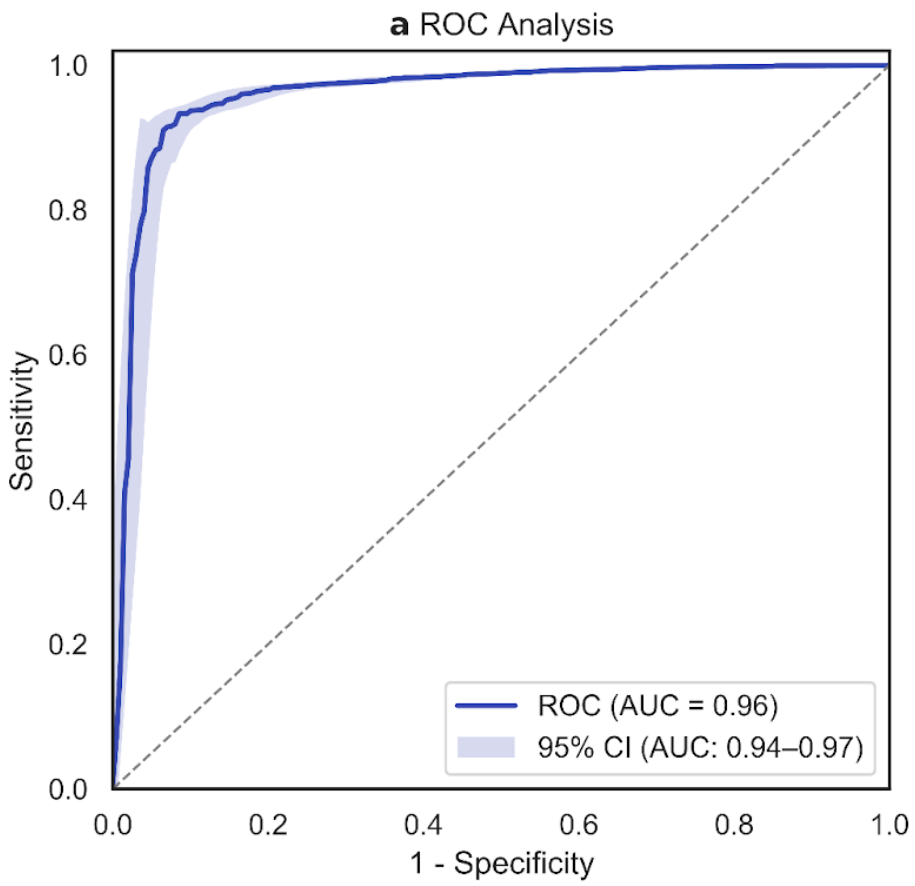
Vilnius University

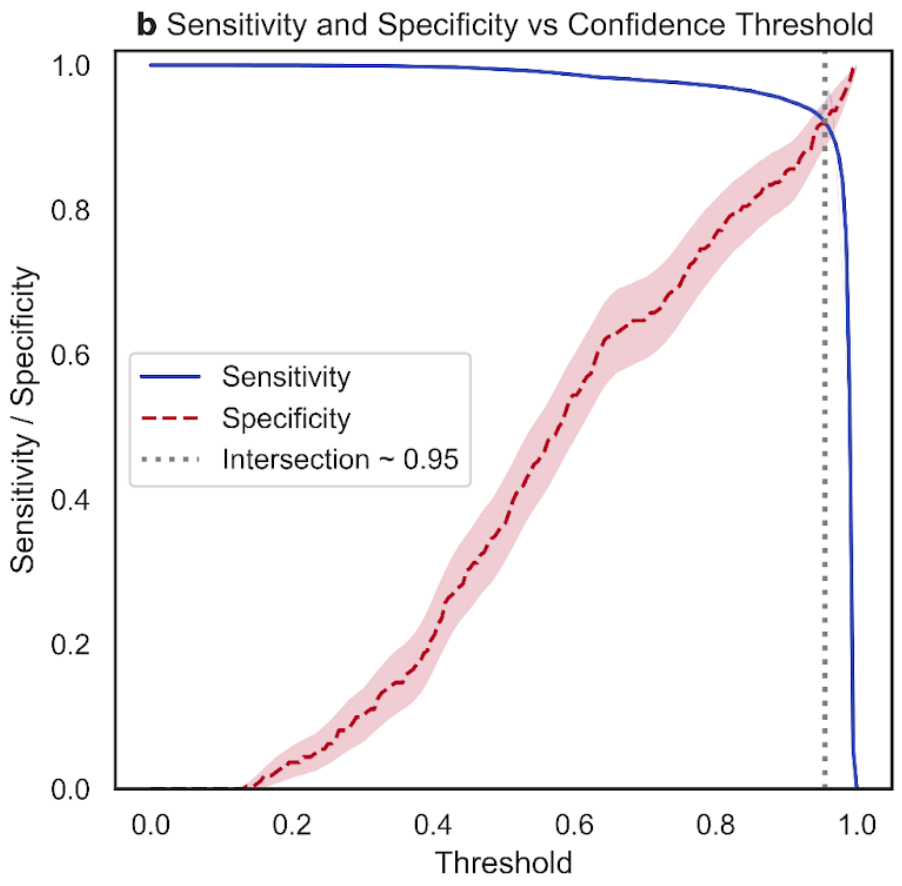
FRANK vs M&M



nius university

ROC analizė (*n*=16,398, tumor and normal)





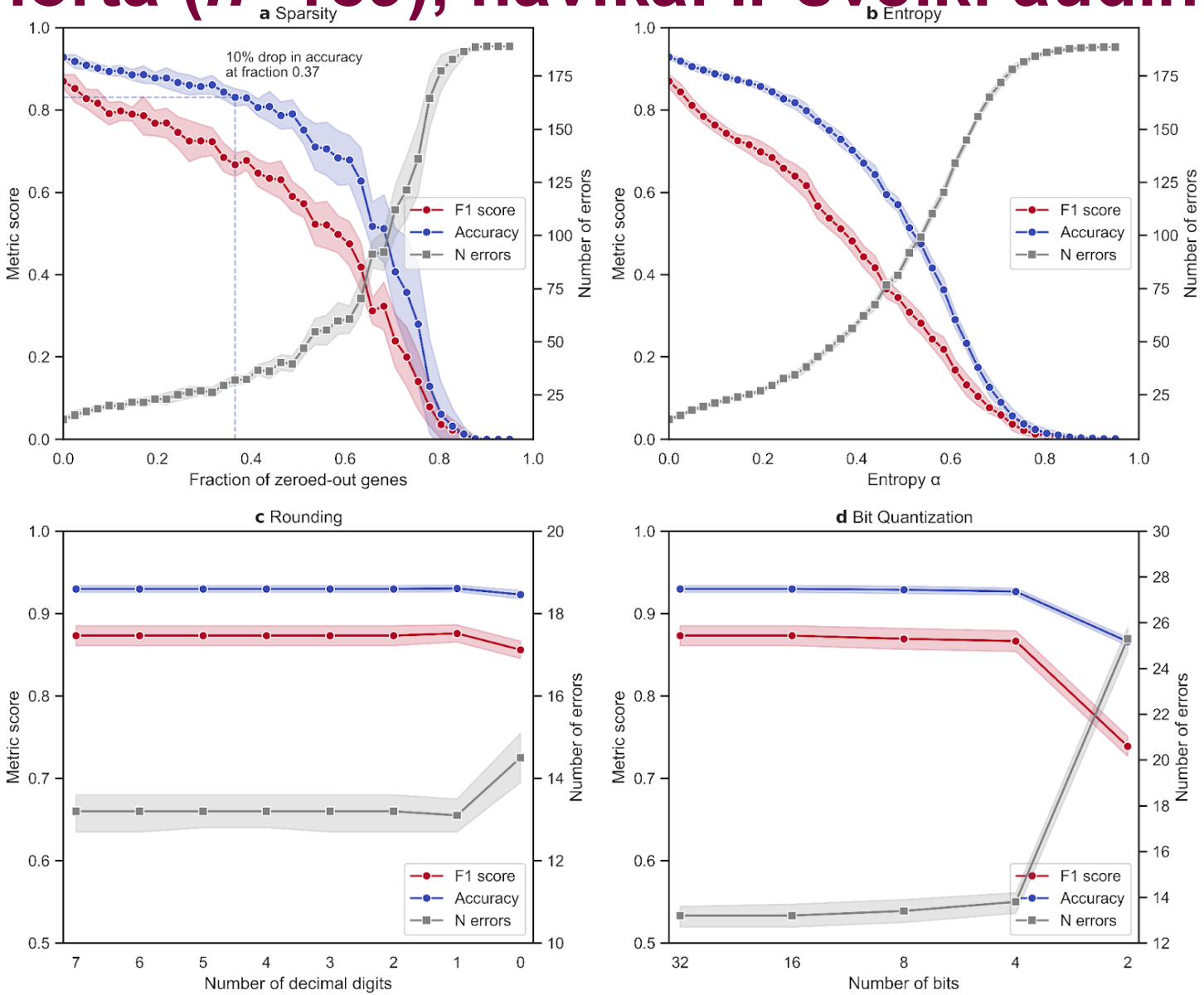
Vilnius University

Vilnius University

Aggregated metrics on scorable samples

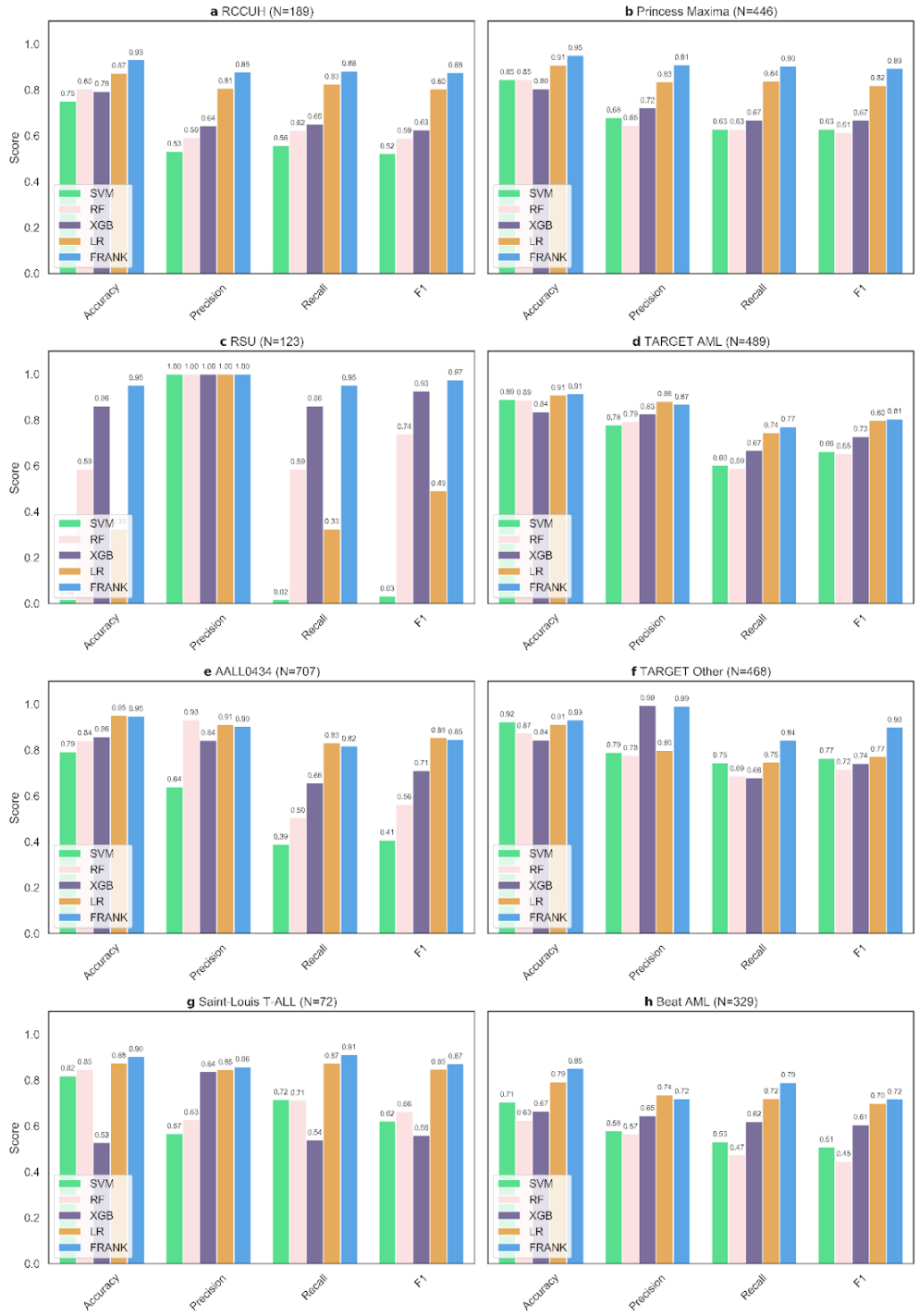
Sample selection	N	N errors	N tumor/normal mismatch	Accuracy (%)	F1 score	Precision	Recall
Tumor and normal tissue - no confidence threshold	16398	272	26	98,3	0,86	0,87	0,88
Tumor and normal tissue - high confidence	14878	22	0	99,9	0,99	0,99	0,99
Tumor only - no confidence threshold	2673	201	18	92,5	0,82	0,84	0,85
Tumor only - high confidence	1785	11	0	99,4	0,98	0,98	0,98
Normal tissue only - no confidence threshold	13725	71	8	99,5	0,98	0,99	0,97
Normal tissue only - high confidence	13093	11	0	99,9	0,99	1	0,99

Modelio tvirtumo (robustness) analizė, RCCUH kohorta (*n*=189), navikai ir sveiki audiniai



Palyginimas su kitais klasifikavimo metodais

- Logistinė regresija - panašiausi rezultatai

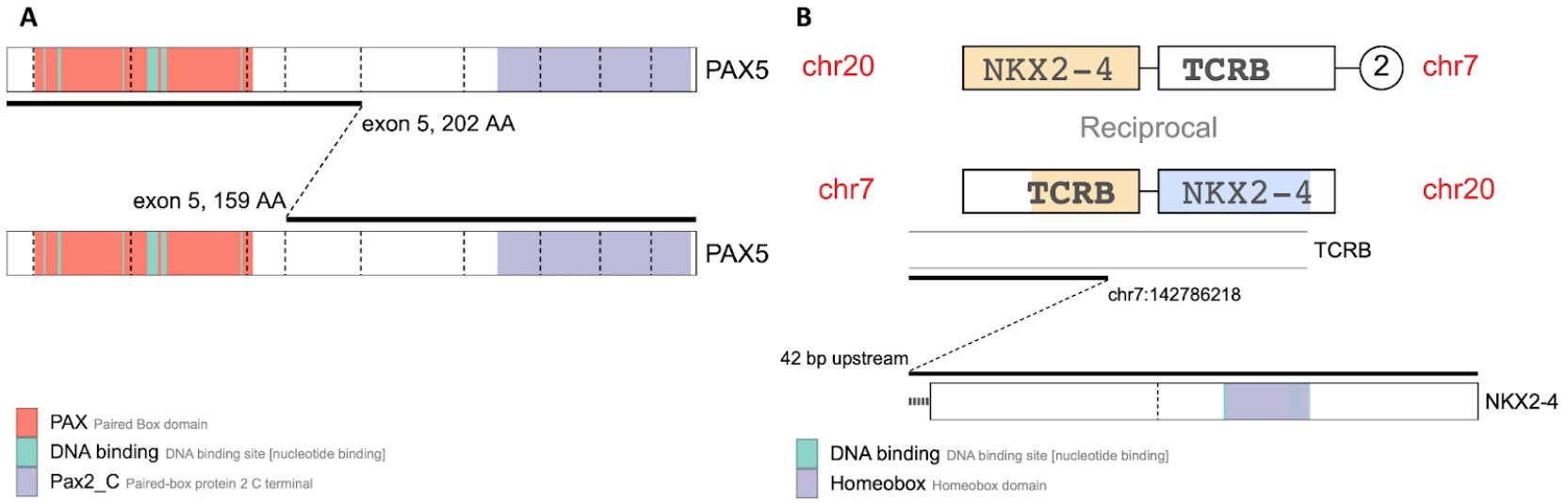


dataset	sample_id	ground_truth	fine_top1_preds	fine_top1_confs
BKUS	22013545RN	T-cell Acute Lymphoblastic Leukemia, NKX2 Activation	T-cell Acute Lymphoblastic Leukemia, NKX2 Activation	0,856143594

Initially labeled as “T-ALL, NOS”

dataset	sample_id	ground_truth	fine_top1_preds	fine_top1_confs
BKUS	23013768DN	B-cell Acute Lymphoblastic Leukemia, PAX5 Alteration	B-cell Acute Lymphoblastic Leukemia, PAX5 Alteration	0,981337

Initially labeled as “B-ALL, NOS”, confirmed to have PAX5 duplication



Vilnius University

Naujumas

Vilnius University

- Aptinkamų navikų skaičius (181 vs 96 prieš tai aprašyti literatūroje).
- Duomenų augmentavimo strategija, CutMix panaudojimas, hierachinės klaidos nuostolio funkcija.
- Išsamus testavimas (M&M – pagrindinė testavimo kohorta iš tos pačios institucijos kaip ir mokymo duomenys).

Kito semestro planas

Vilnius University

- Publikuotas pirmas rankraštis (Science Translational Medicine arba Communications Medicine žurnalai)
- Pradėtas antras rankraštis
- Pranešimo skaitymas tarptautinėje konferencijoje (European Society of Human Genetics)