



**Vilnius
universitetas**

Doktorantas:
Brendonas Stakauskas
2022-2026

Darbo vadovas:
Dr. Virginijus Marcinkevičius

2024/2025 metai
III metai, I pusmetis



**Giliais neuroniniais tinklais grįstų
mašininio mokymo metodų taikymas
viruso mutacijų trajektorijai
prognozuoti**

TURINYS

1. Problemos apibrėžimas, tyrimo objektas ir tikslai
2. Studijų plano vykdymas
3. Trumpas per pusmetį gautų mokslinių rezultatų pristatymas
4. Kito pusmečio darbo planas



**Problemos
apibrėžimas,
tyrimo objektas
ir tikslai**

Tyrimo objektas

Virusų baltymų sekos ir giliaisiais neuroniniais tinklais grįsti mašininio mokymo algoritmai skirti prognozuoti viruso mutacijas.

Tyrimo problemos

- Sąryšio tarp viruso proteinų sekų nustatymas;
- Duomenų aibės, atspindinčios istorines mutacijas, sudarymas;

Tyrimo tikslas

Sukurti giliaisiais neuroniniais tinklais ir natūralios kalbos apdorojimo algoritmais grįstą metodą leidžianti numatyti viruso mutacijos trajektoriją.

Tyrimo uždaviniai

- Atlikti literatūros analizę, išanalizuoti *state-of-the-art* algoritmus viruso baltymo mutacijų prognozavimui.
- Atkartoti literatūroje pateikiamų metodų rezultatus.
- Sukurti metodą duomenų aibės, atspindinčios istorines mutacijas, sudarymui.
- Sukurti duomenų aibę tyrimui.
- Pasiūlyti naują metodą viruso mutacijoms numatyti.
- Atlikti eksperimentinius tyrimus, palyginant pasiūlytą metodą su literatūroje aprašytais metodais.

Studijų plano vykdymas

Studijų planas, vykdymo suvestinė

Studijų metai	Egzaminai		Dalyvavimas konferencijose				Publikacijos				
			Nacionalinėse		Tarptautinėse		Su citavimo rodikliu		Be citavimo rodiklio		Buklė
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	
I (2022/2023)	2	2	1	0	0	1	0	0	1	1 (konferencijos darbų medžiagoje)	
II (2023/2024)	2	2	1	0	0	1	0	0	1	0	
III (2024/2025)	0	0	0	1	1	0	1	0	0	0	
IV (2025/2026)	0	0	0	0	1	0	1	0	0	0	
Iš viso:	4	4	2	1	2	2	2	0	2	1	

Ataskaitinio pusmečio darbo planas ir jo įvykdymas

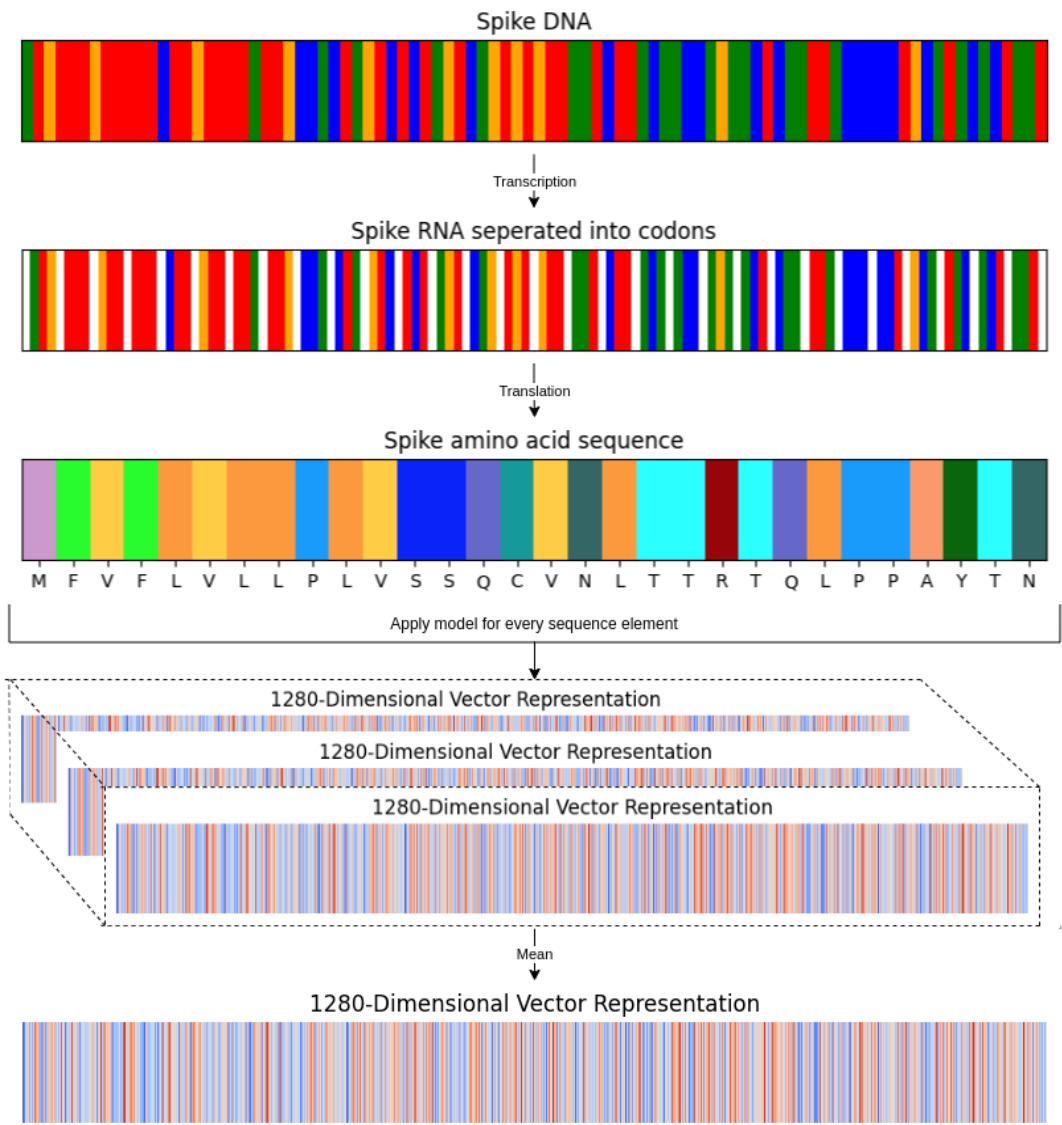
Dalyvavimas konferencijose 2024/2025 (1 pusmetis)

Planas	Įvykdyta	Konferencijos tipas
Data Analysis Methods for Software Systems, 2024 Lapkričio 28-30, Druskininkai, Lietuva	Poster B. Stakauskas "Visualising SARS-CoV-2 Phylogenetic Relationships Using Protein Language Models "	Nacionalinė

Visų mokslinių tyrimų ir disertacijos rengimo etapai

1. Mokslinių tyrimų disertacijos tema apžvalga ir analizė
2. Mokslinio tyrimo vykdymas
 1. Tyrimo metodikos sudarymas
 2. Teorinis tyrimas
 3. Empirinis tyrimas
 4. Gautų duomenų analizė, apibendrinimas, išvadų parengimas
3. Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas
4. Daktaro disertacijos parengimas ir svarstymas padalinyje
5. Daktaro disertacijos gynimas

Trumpas per pusmetį gautų mokslinių rezultatų pristatymas



DNR transformacijos į įterpinį proceso iliustracija

Įterpinių palyginimas

- Naudojami ESM-2 ir ProtT5 modelių įterpiniai
- Sumažinama duomenų dimensija
- Naudojamas klasterizavimo metodas HDBSCAN sudarant klasterius
- Klasteriai palyginami su priskirtais kladais

Duomenų rinkinys

- 1523 H3N2 viruso sekų
- Hemaglutinino genomas (567 amino rūgščių ilgio)
- Duomenys paimti iš straipsnio *Dimensionality reduction distills complex evolutionary relationships in seasonal influenza and SARS-CoV-2*

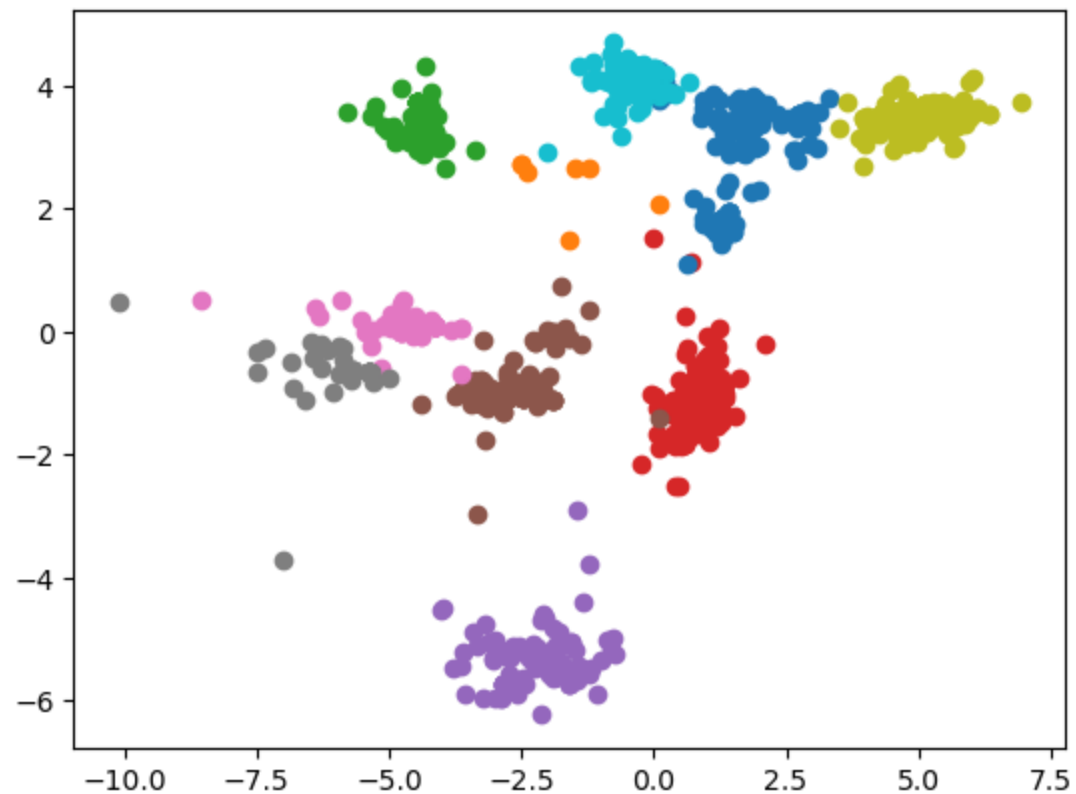
Modeliai

- ESM-2
 - t-33_650M 1280 dimensijų įterpinys
 - t-36_3N 2560 dimensijų įterpinys
 - t-48_15B 5120 dimensijų įterpinys
- ProtT5-XL 1024 dimensijų įterpinys

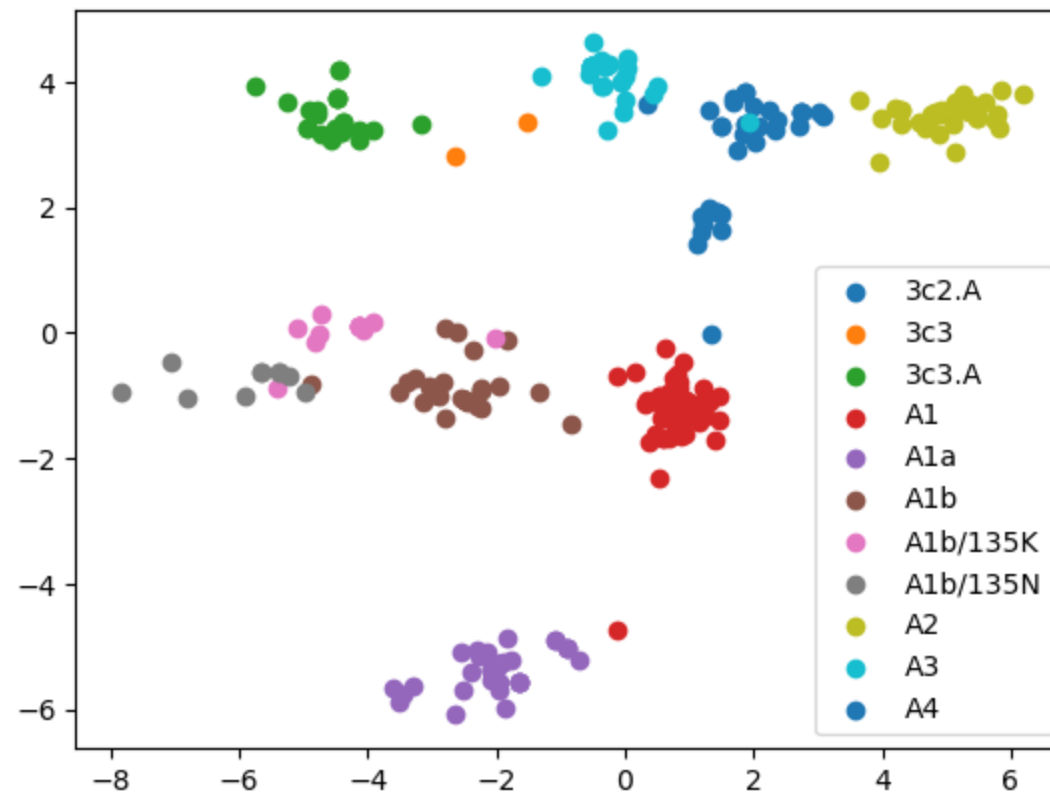
Dimensijos mažinimo metodai

- t-sne
- UMAP
- Triplet Network (WIP)

Triplet Network



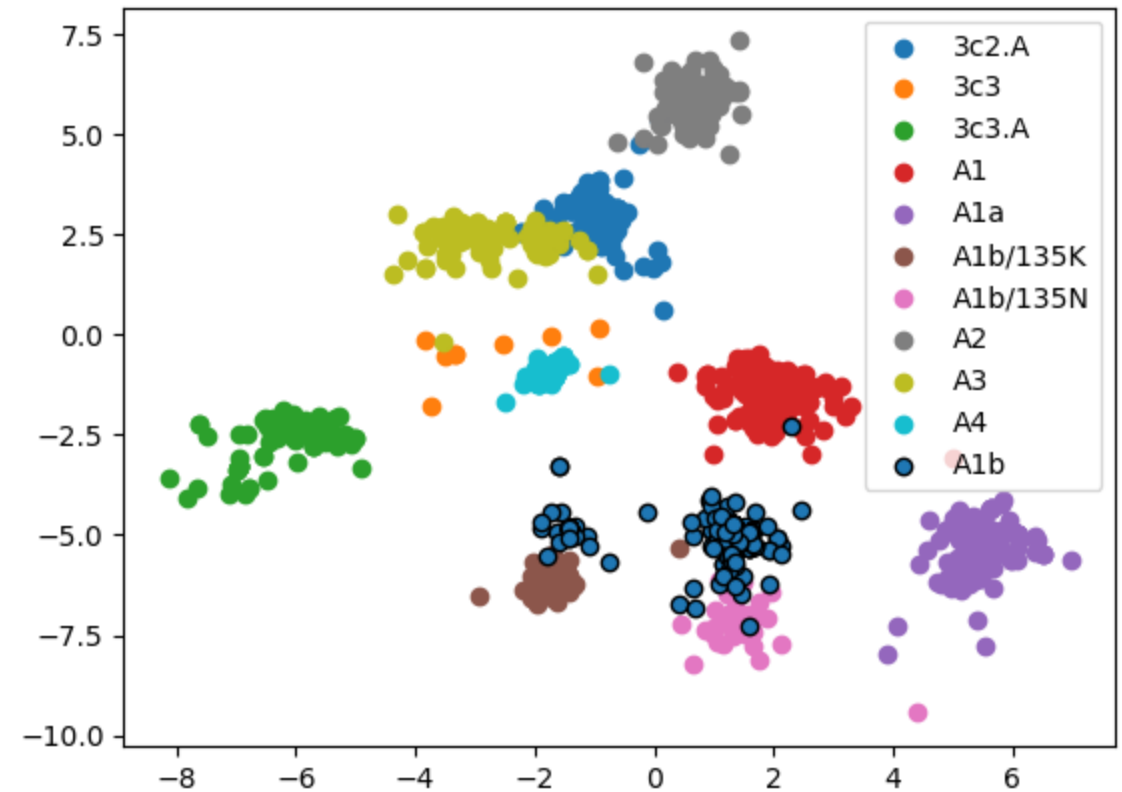
Train



Test

Triplet Network (2)

Testavimo aibė paremta duomenimis iš vieno nematyto klado. Persimaišo su esamais kladais, sunku išskirti kaip naujus duomenis.



Rezultatai

Metodas	Normalizuotas VI
t-SNE	0.08
UMAP	0.09
PCA	0.11
ESM2-t36 + t-SNE	0.11
ESM2-t48 + t-SNE	0.12
ESM2-t33 + t-SNE	0.14
ProtT5 + t-SNE	0.16

Triplet network rezultatai dar nėra galutiniai, todėl neįtraukti

Kito pusmečio darbo planas

Mokslinė veikla:

- Atrinktų algoritmų įgyvendinimas, modifikavimas ar naujų algoritmų kūrimas, sprendžiant apibrėžtus uždavinius.
- Eksperimentinis sukurtų modifikacijų tyrimas, analizuojant jų efektyvumą.

Dalyvavimas konferencijose

Dalyvavimas konferencijoje Suomijoje ISBRA2025

Dalyvavimas konferencijoje Kaune ICANN2025

Publikacijų rengimas

Publikacijos rengimas Computational Biology, BMC Bioinformatics ar panašiam leidinyje turinčiame Impact Factor.