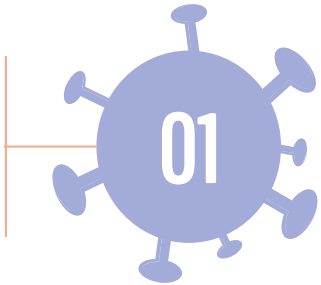
"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER
SCIENCE AND ENGINEERING

Language model that semantically captures masked glycoprotein amino-acid mutations in emerging variants of the SARS-CoV-2 virus

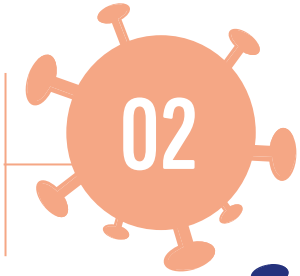
Mentor: Phd Prof. Nevena Ackovska
Student: Kiril Zelenkovski

PROJECT CONTENT



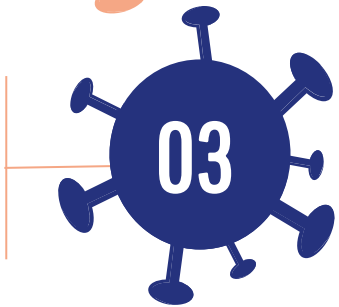
1. Introduction

- SARS-CoV-2 emergence
- Importance of the glycoprotein



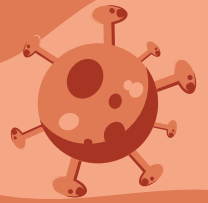
2. Dataset

- Data interpretation
- Data acquisition [GISAID]
- Data preprocessing
- Data analysis



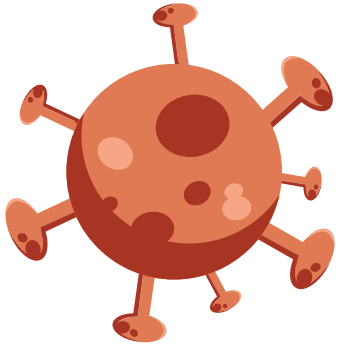
3. Models, Reults

- Transformers architecture
- Results and discussion



1. Introduction

- How did SARS-CoV-2 happen?
- What does the anatomy look like?
- How important is the S-gene?



SARS-CoV-2

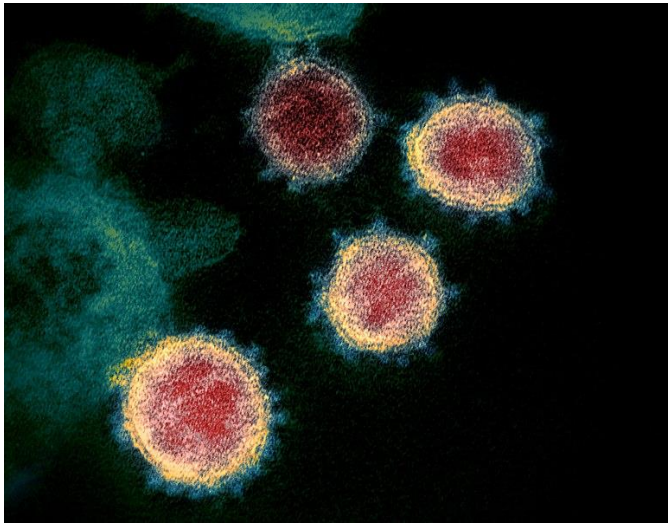


Fig1. This transmission electron microscope image shows SARS-CoV-2

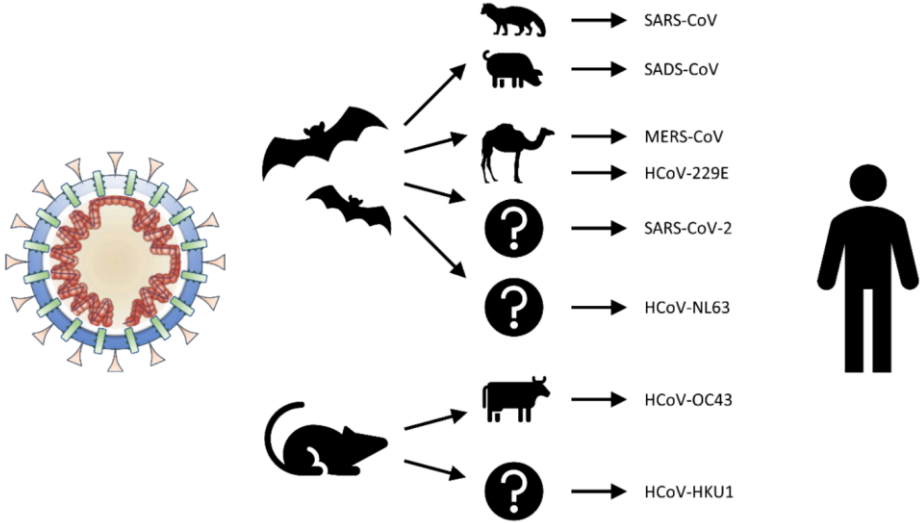


Fig2. Zootonic spillover, possible intermediate hosts

Anatomy of the SARS-CoV-2

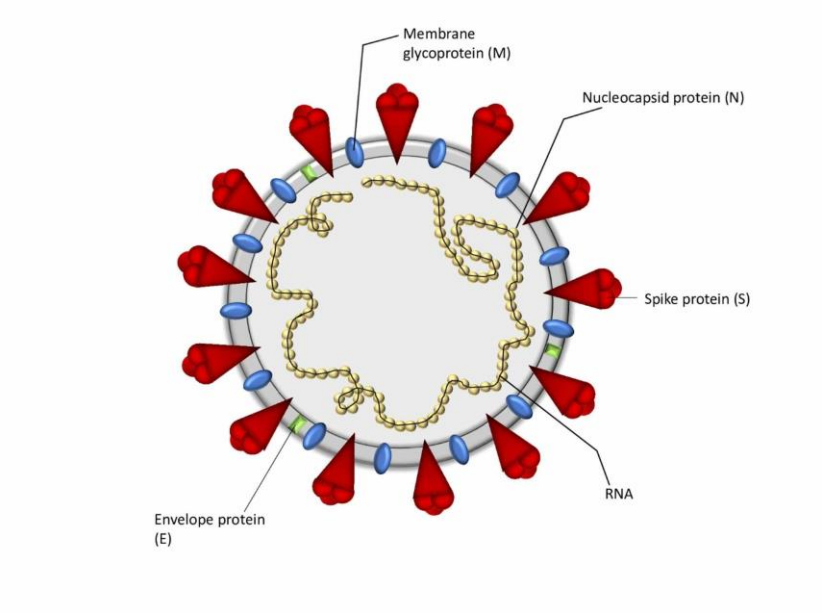


Fig3. Viral particle of the SARS-CoV-2

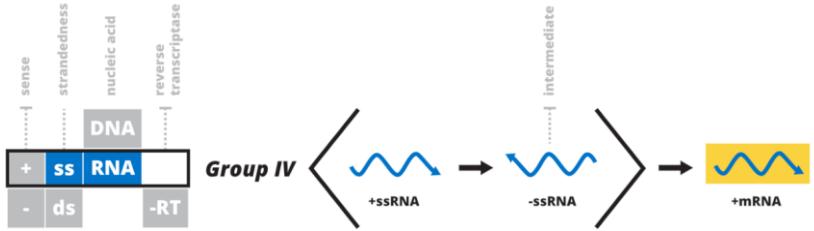


Fig4. Baltimore classification for the positive sense single-stranded RNA viruses

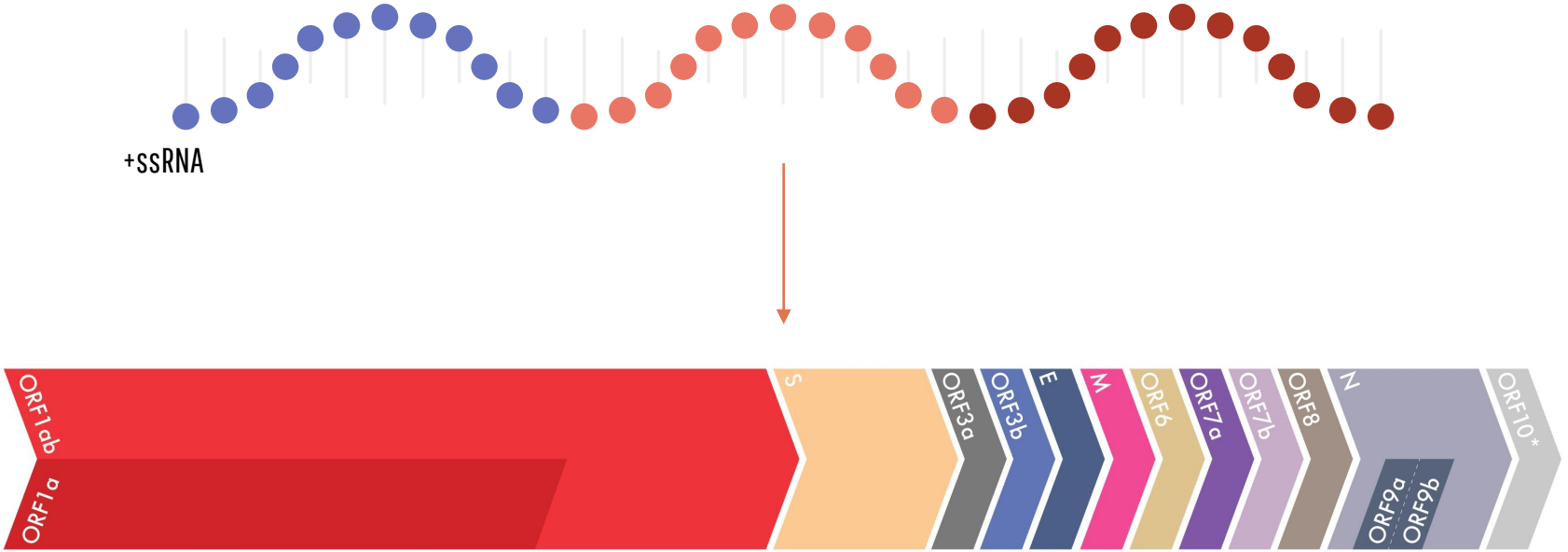


Fig5. SARS-CoV-2 isolate chart Wuhan-Hu-1;
complete genome

type: CDS
location: [21562:25384](+)
qualifiers:
    Key: codon_start, Value: ['1']
    Key: db_xref, Value: ['GeneID:43740568']
    Key: gene, Value: ['S']
    Key: gene_synonym, Value: ['spike glycoprotein']
    Key: locus_tag, Value: ['GU280_gp02']
    Key: note, Value: ['structural protein; spike protein']
    Key: product, Value: ['surface glycoprotein']
    Key: protein_id, Value: ['YP_009724390.1']
    Key: translation, Value: ['MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIF
GTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLA
LHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCY
GVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFEL
LHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAE
HVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT
PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFN
SAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHV
TYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQE
LGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT']

SARS-CoV-2 binding cites

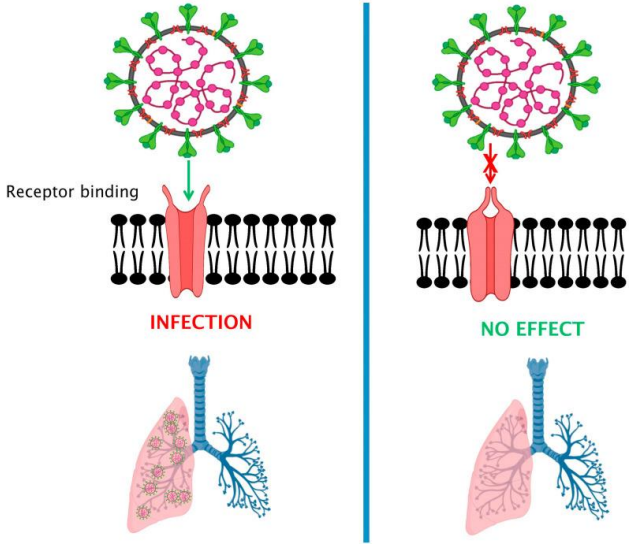
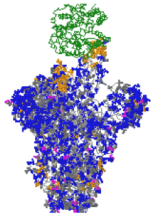
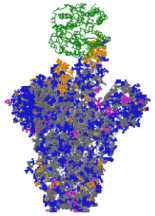


Fig5. Different interactions between CoVs and their cellular receptors[source]

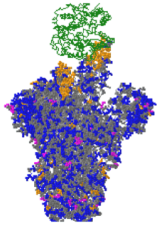
N501S K417N E484Q V503F Y453F
F490Y E484A S477G N439K S494L
S477R F490S F456L G446V Q493K
S477N S477I G476S N501T T478R
S494A L455F N501Y G446S E484K
Y449N T478K Q493L S494P
in GH clade (B.1.*)



N501S G496V V445F E484Q T478I
Y449H S494L G446V S477N F490L F486L
N501T G476S T478R N501Y G446L
V503A E484K K458N Q493L G485R
T478K T500S S494P F456Y Q493R
K417N V503F A475S E484A N439K
Q493E S477G E484R Y449S A475V
F490S S477I G476D P499L V503L K417T
G485D R403K V445I L455F G485S G446S
P499S R403I
in G & GV clades (B.1 & B.1.177)



E484Q T478I F490I G504S S477N Y453E
G476S F486I Y505H K417M N501Y E484K
K417R S477T K458N Q493L E484G V503F
G504D Y453F Y449S G485V Y489H K417T
R403K G485S G504N G446S F486S V445F
P499H V445A G496R Y449H S494L Q493H
F456L G446V F490L N501T T478R G496S
G446A G496W Y495I G447S Y449N T478K
G485R S494P K417A F456Y Q493R K417N
N501I R403P A475S N439K A475V S477R
F490S Q493K S477I V503L L455F V445I
E484D G485A R403I
in GR & GRY clade (B.1.1.1 & B.1.1.7)



N501T E484Q K417N T478R K417T
N501Y E484K T478K Y449H S494P
S477R F490S S477N F490L
in nonG clade (A, B & B.2)

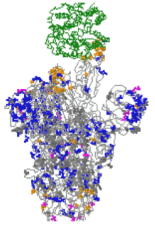
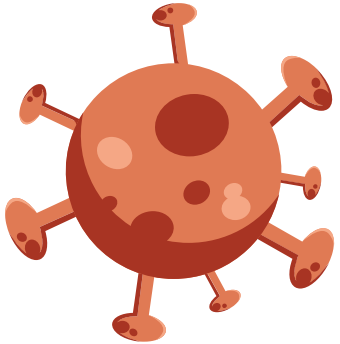


Fig6. Receptor binding surveillance for complete genomes [GISAID]

2. Dataset

- Data interpretation: Lineages / Variants
- Data acquisition pipeline[GISAID]
- Data preprocessing
- Data analysis



Data interpretation

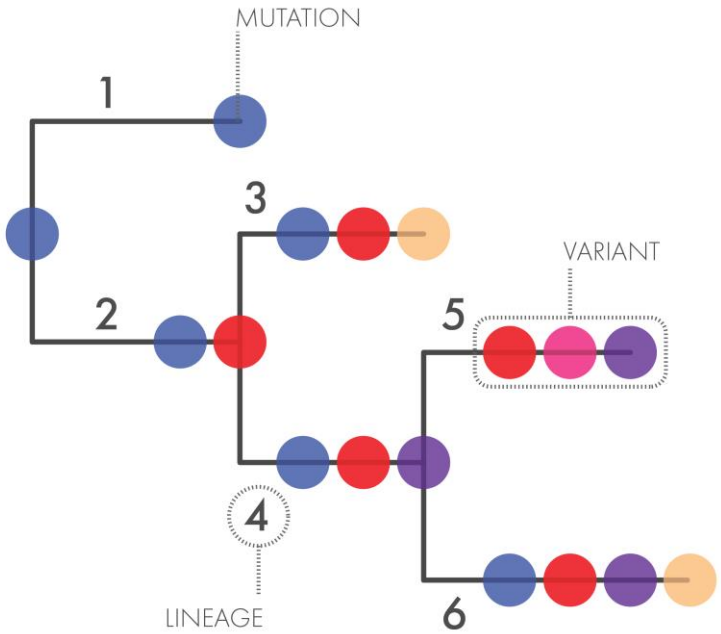


Fig7. Glossary terms for better data interpretation

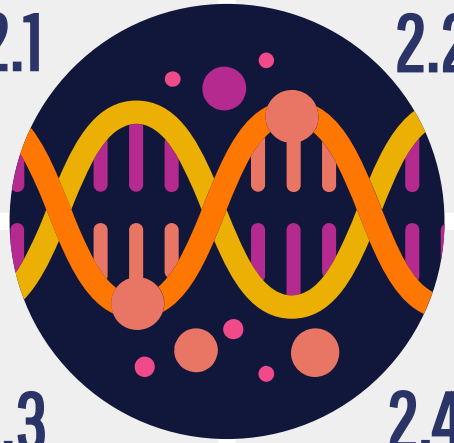
Data acquisition pipeline

Read emerging variants [July 2021 – June 2021]

2.1

2.2

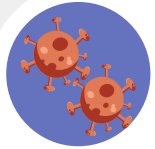
Extract all lineages, locate their mutations



Create peptides from lineage/variant data

2.3

2.4



Clean, analyse and create training csv

hosted by the Federal Republic of Germany

developed by

max planck institut informatik

2.1. Read emerging variants

Variant	#Genomes	#Top Location	#Top Clade	#Top Lineage	Co-occurring Changes List	#Co-occurring Changes	Δ#Loc(S)	#aachanges(C)	(SxC)
452R_478K_681R	56821	37945 England	56752 G	51938 B.1.617	Spike_T19R, Spike_E156G, S	24	16	3	48
69X_70X_452R_478	79	17 England	78 G	71 B.1.617.2	Spike_T19R, Spike_E156G, S	16	2	5	10
144del_452R_478K	28	4 Delhi	28 G	23 B.1.617.2	Spike_D614G, Spike_T19R,	11	2	4	8
417N_452R_478K_6	204	63 California	204 G	181 AY.1	Spike_T19R, Spike_E156G, S	16	2	4	8
367L_452R_478K_6	7	3 England	7 G	5 B.1.617.2	Spike_E156G, Spike_D950N	15	2	4	8
244X_452R_478K_6	19	6 Ontario	19 G	18 B.1.617.2	Spike_D614G, Spike_D950N	12	2	4	8
452R_478K_490L_6	8	3 Scotland	8 G	7 B.1.617.2	Spike_D614G, Spike_D950N	10	2	4	8
5F_69del_70del_14	3	1 England	2 G	2 B.1.617.2	Spike_D614G, Spike_Y145d	4	1	7	7
242X_243X_244X_4	5	1 Doha	5 G	4 B.1.617.1	Spike_Q1071H, Spike_D614	11	1	6	6
5F_69del_70del_14	6	3 Norrbotten	6 GRY	6 B.1.1.7	Spike_A570D, Spike_Y145d	21	1	6	6
18F_417T_484K_50	139	29 California	116 GR	112 P.1	Spike_D138Y, Spike_R190S,	21	1	5	5
18F_69del_70del_1	22	8 Arizona	22 GR	21 B.1.1.7	Spike_A570D, Spike_P681H	19	1	5	5

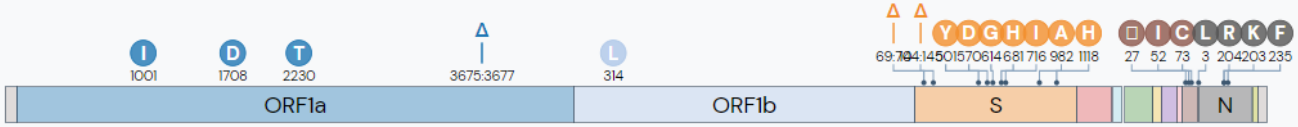
Fig8. First rows of emerging variants table for July 21'

Characteristic mutations of B.1.1.7

gene	amino acid
ORF1a	T1001I
ORF1a	A1708D
ORF1a	I2230T
ORF1a	del3675/3677
ORF1b	P314L
S	del69/70
S	del144/145
S	N501Y
S	A570D
S	D614G
S	P681H
S	T716I
S	S982A
S	D1118H
ORF8	Q27*
ORF8	R52I
ORF8	Y73C
N	D3L
N	R203K
N	G204R
N	S235F

outbreak.info

2.2: Extract all Lineages and find metadata



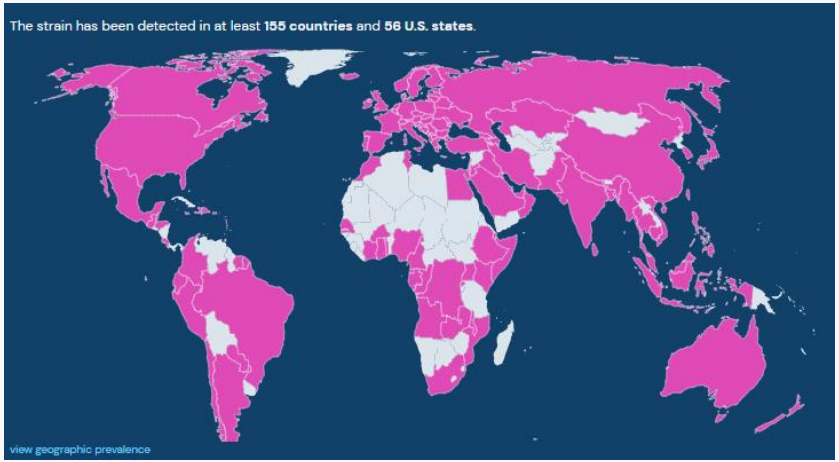


Fig9. B.1.1.7 Lineage Report and table of mutations

2.3. Create peptides

```
B.1.1.523
Mutation is of type - substitution  : T 1027 I
Mutated T into I
Mutation is of type - substitution  : D 614 G
Mutated D into G
Mutation is of type - substitution  : E 484 K
Mutated E into K
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGV
```

```
69X_439K
Relevant aa changes / Antibody cites:  ['69X', '439K']
Co-occurring changes (only in Spike):  ['T19R', 'E156G', 'D950N', 'F157del', 'R158del', 'D614G']
['F157del', '439K', '69X', 'E156G', 'R158del', 'D614G', 'T19R', 'D950N']
Mutating  N into K
Mutating  H into X
Mutating  E into G
Mutating  D into G
Mutating  T into R
Mutating  D into N
MFVFLVLLPLVSSQC...T
```

Fig10. Example outputs of peptide creation for random lineage (up), variant (down)

2.4. Cleaning and analyzing data

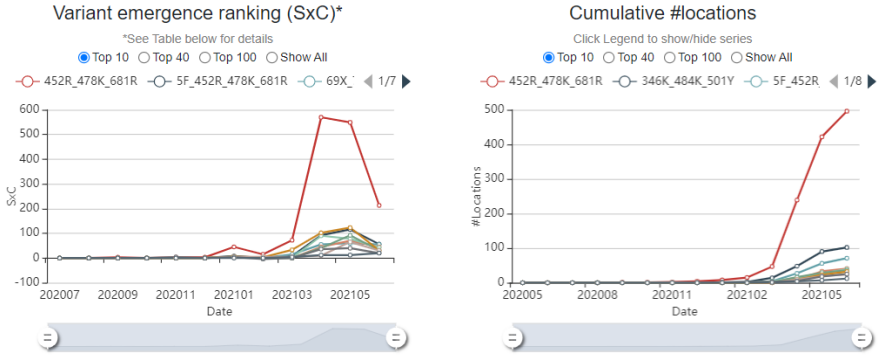


Fig11. GISAID plots on the emergence ranking and number of locations

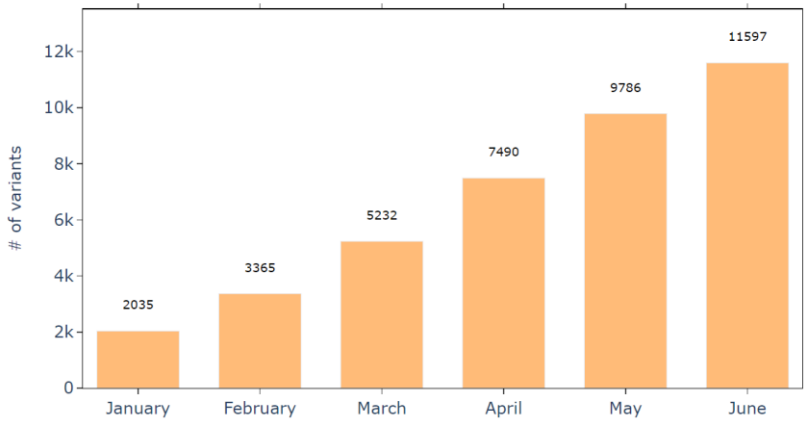
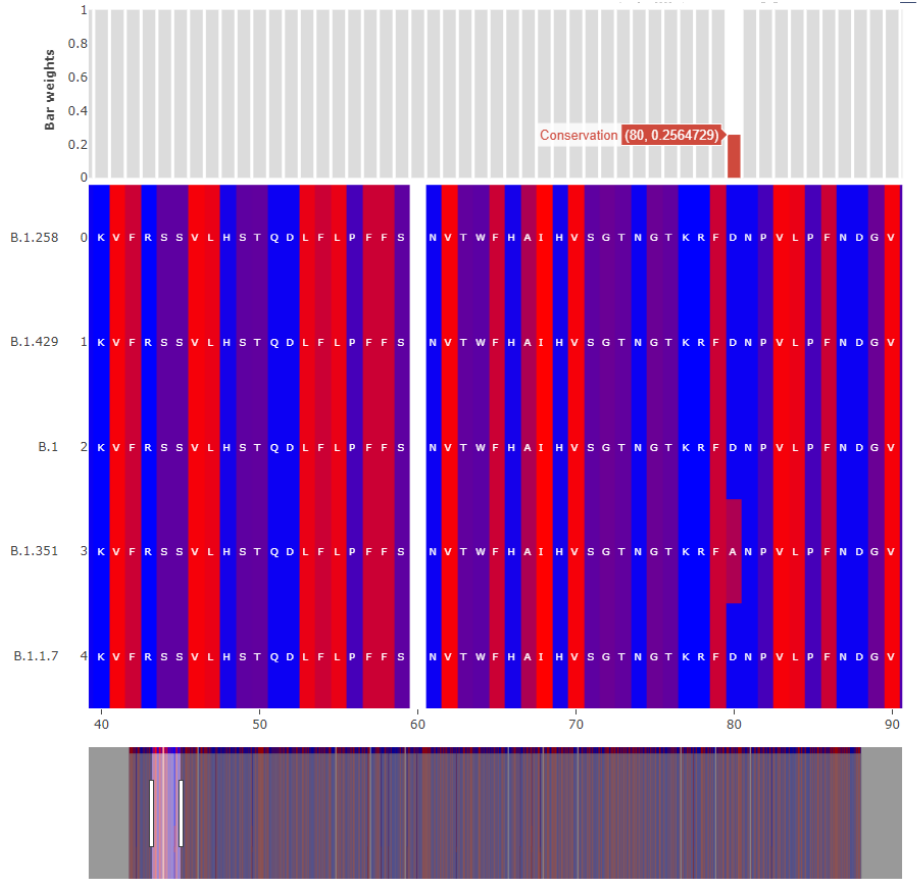


Fig12. Number of variants per month



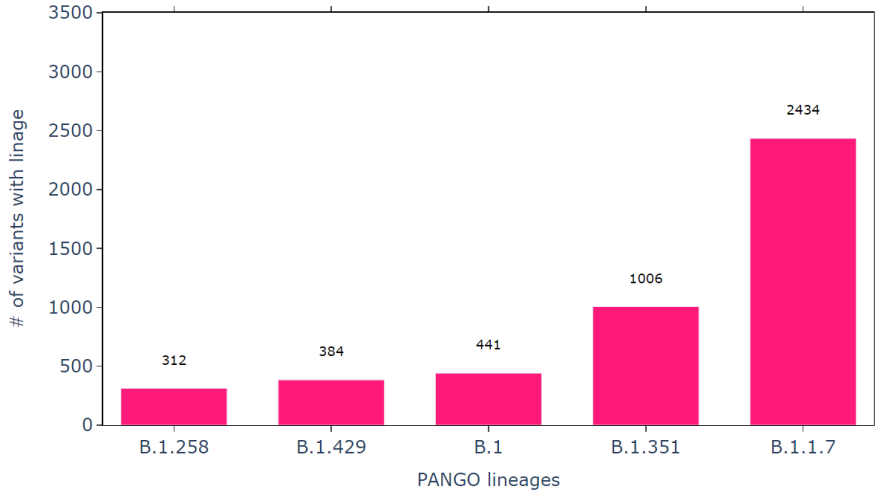
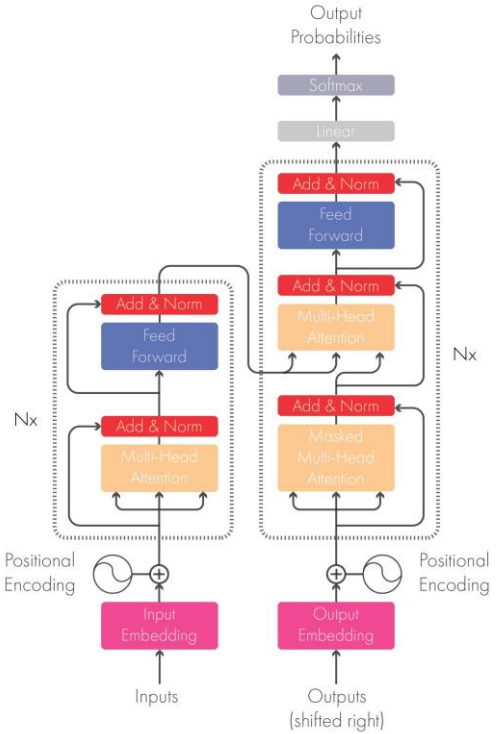


Fig13. Alignment chart and bar plot of the 5 most common lineages

3. Models, Results

- What are Transformers?
- Attention is all
- Training and evaluation
- Masking mutations

“Attention is all you need”



$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Fig14. The classical Transformer architecture; diagram (left); formula (right)

Idea [Esperanto]

```python
from transformers import pipeline

fill_mask = pipeline(
    "fill-mask",
    model="./models/EsperBERTo-small",
    tokenizer="./models/EsperBERTo-small"
)

# The sun <mask>.
# =>

result = fill_mask("La suno <mask>.")

# {'score': 0.2526160776615143, 'sequence': '<s> La suno brilis.</s>', 'token'
# {'score': 0.0999930202960968, 'sequence': '<s> La suno lumis.</s>', 'token':
# {'score': 0.04382849484682083, 'sequence': '<s> La suno brilas.</s>', 'token
# {'score': 0.026011141017079353, 'sequence': '<s> La suno falas.</s>', 'token
# {'score': 0.016859788447618484, 'sequence': '<s> La suno pasis.</s>', 'token
```

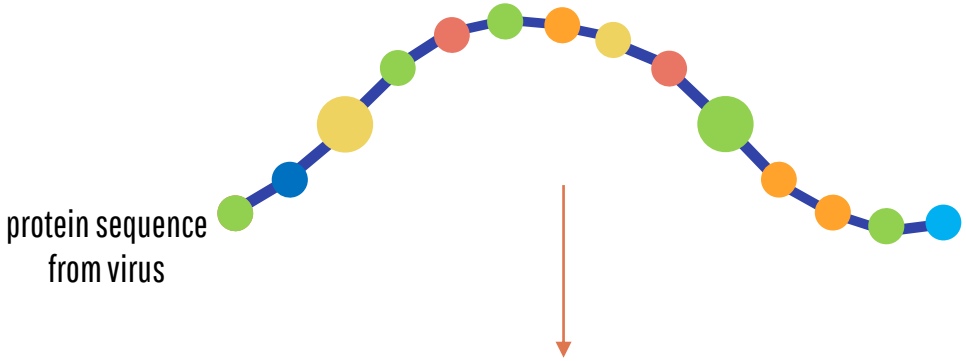
```python
fill_mask("Jen la komenco de bela <mask>.")

# This is the beginning of a beautiful <mask>.
# =>

# {
#     'score':0.06502299010753632
#     'sequence':'<s> Jen la komenco de bela vivo.</s>'
#     'token':1099
# }
# {
#     'score':0.0421181358397007
#     'sequence':'<s> Jen la komenco de bela vespero.</s>'
#     'token':5100
# }
# {
#     'score':0.024884626269340515
#     'sequence':'<s> Jen la komenco de bela laboro.</s>'
#     'token':1570
# }
# {
#     'score':0.02324388362467289
#     'sequence':'<s> Jen la komenco de bela tago.</s>'
#     'token':1688
# }
# {
#     'score':0.020378097891807556
#     'sequence':'<s> Jen la komenco de bela festo.</s>'
#     'token':4580
# }
```

Fig15. **FillMaskPipeline**
simple sentence (left), complex sentence (right)

Tokenization




```
[ ]  1 print(my_lineages[12])
     2 x = my_lineages[12]
     3 print(x[-1])
     4 print(tokenizer.encode(x).tokens[-2])
     5 print(tokenizer.encode(x))
     6 tokenizer.encode(x).tokens[-1]
```

```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVS
T
T
Encoding(num_tokens=1275, attributes=[ids, type_ids, tokens, offsets, attention_mask, special_tokens_mask, overflowing])
```

Fig16. Tokenization for random peptide lineage

Training [Phase 3]

TL

Tze Chuen LEE <leetc@bii.a-star.edu.sg>
Sun 6/20/2021 6:53 PM
To: Кирил Зеленковски
Cc: GISAID Support <service@gisaid.org>

Dear Kiril,

Thank you for your email and well wishes. We are now working on getting all the co-occurring changes to be included in the variantswatchlist.csv download. We will keep you updated again on its progress.

Currently, we monitor 147 amino acid changes and deletions in the Spike protein that occur in at least 10 different geographical locations and were identified in studies to cause antibody escape, increase ACE2 binding or increase Spike protein expression and stability are considered as part of combinations or constellations forming potential variants to be monitored. These 147 aa changes are the ones found in the "Variant" column. Please note that we constantly search the literature for new mutations that causes similar phenotypic changes so this watch list of amino acid changes may change (increase) over time. For all other amino acid changes that co-occur in >75% of all isolates with the variant (combinations of mutations) in the "Variant" column, we list them in the "Co-occurring changes" column. The list of mutations found in the co-occurring changes column may still be interesting to researchers who are tracking other possible contributing characteristics to these variants.

We hope this answer your questions.

Best regards,
Raphael (on behalf of GISAID EpiCoV Team)

From: <kiril.zelenkovski@students.finki.ukim.mk>
Date: Sun, Jun 20, 2021 at 7:19 AM
Subject: Missing data from Emerging Variants downloads
To: <service@gisaid.org>

Kirl Zelenkovski writes,

Dear Sir/Madam,

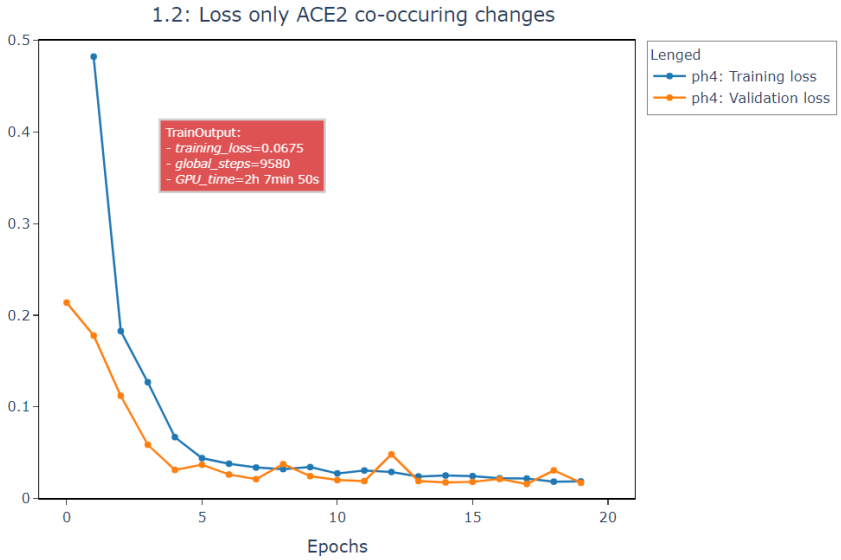
I encounter missing data when downloading data from the Emerging variants option. The "variantswatchlist.csv" seem to miss some of the mutations. For example, when I hover over the data (on the web page) it shows me all of the "Co-occurring changes" (for example if the "#Co-occuring changes" is 16 it shows all of them) but when I download the data only the first 3 are available in the CSV. I am trying to predict those changes and it is very difficult for me to manually write down the remaining mutations for each of the variants for the last 10 months.

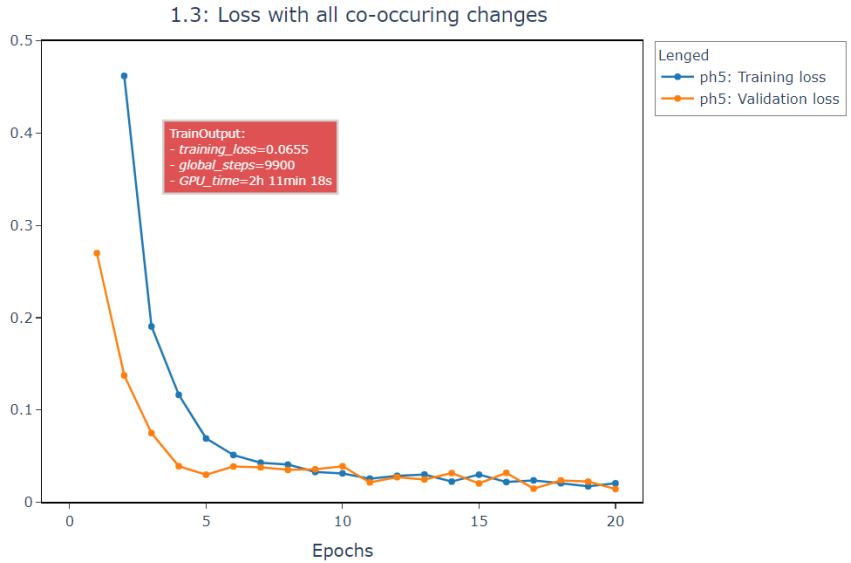
Also, it would mean the world to me if you could explain to me what exactly does "Co-occurring changes" and Variants mean? Why sometimes the relevant changes ("Variant" column) does not include the "Co-occurring changes"?

Can wait for your reply really means a lot. Stay safe.

With respect,
Kiril

Training [Phase 4 vs. Phase 5]





Expected result

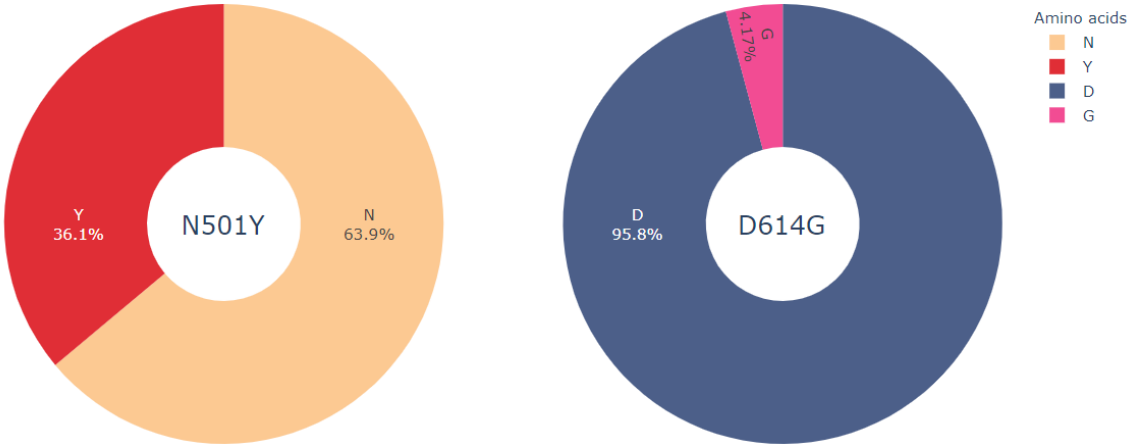


Fig17. Expectations based on frequency analysis of position 501 and 614 in input sentences

```
test = list(my_variants[0])
test[613] = '<mask>'
B_TEMP_masked = "".join(test)
fill_mask(B_TEMP_masked)
```

Captured scores

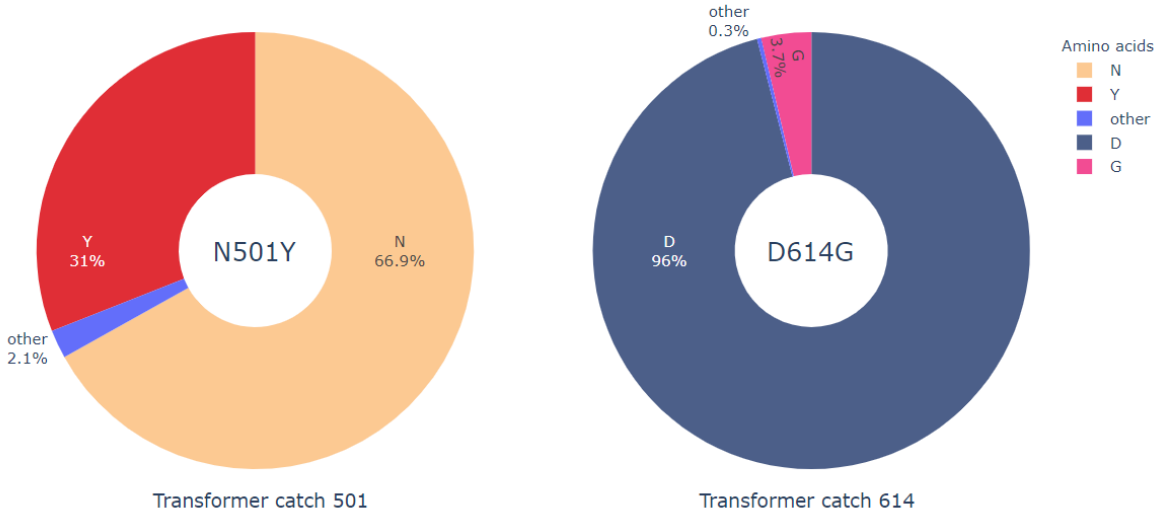
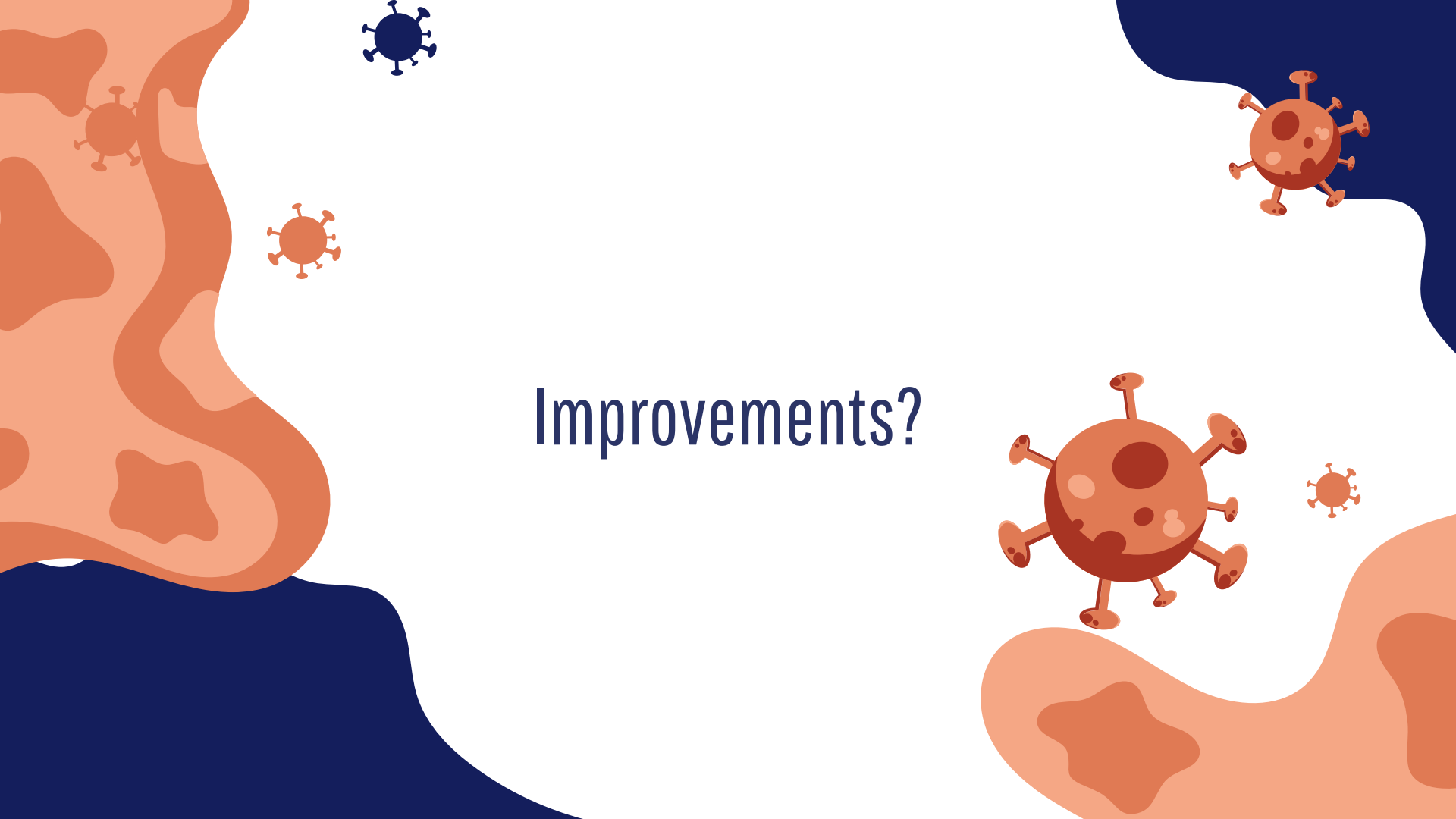


Fig18. Actual scores of the model for mutations N501 and D614G

Improvements?



Acknowledgment

Which animals are being used to develop a COVID-19 vaccine?

MICE

Mice are being used to test whether vaccine compounds are safe to be trialled in humans. There is only one strain of genetically altered mice that is susceptible to COVID-19. These mice were developed to research the SARS outbreak in 2003 and are now being bred for COVID-19 research.

MONKEYS

Non-human primates are our closest living relatives. Unlike mice, they can contract the COVID-19 virus. Researchers are using primates to test the safety of vaccine compounds, discover how the virus works inside the body, and whether it can re-infect people that have already recovered from the virus.

Understanding ANIMAL RESEARCH

"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER
SCIENCE AND ENGINEERING

Discussion

Mentor: Phd Prof. Nevena Ackovska
Student: Kiril Zelenkovski

slidesgo