MACHINE LEARNING CLUSTERING OF SPIKE PROTEIN MUTATIONS

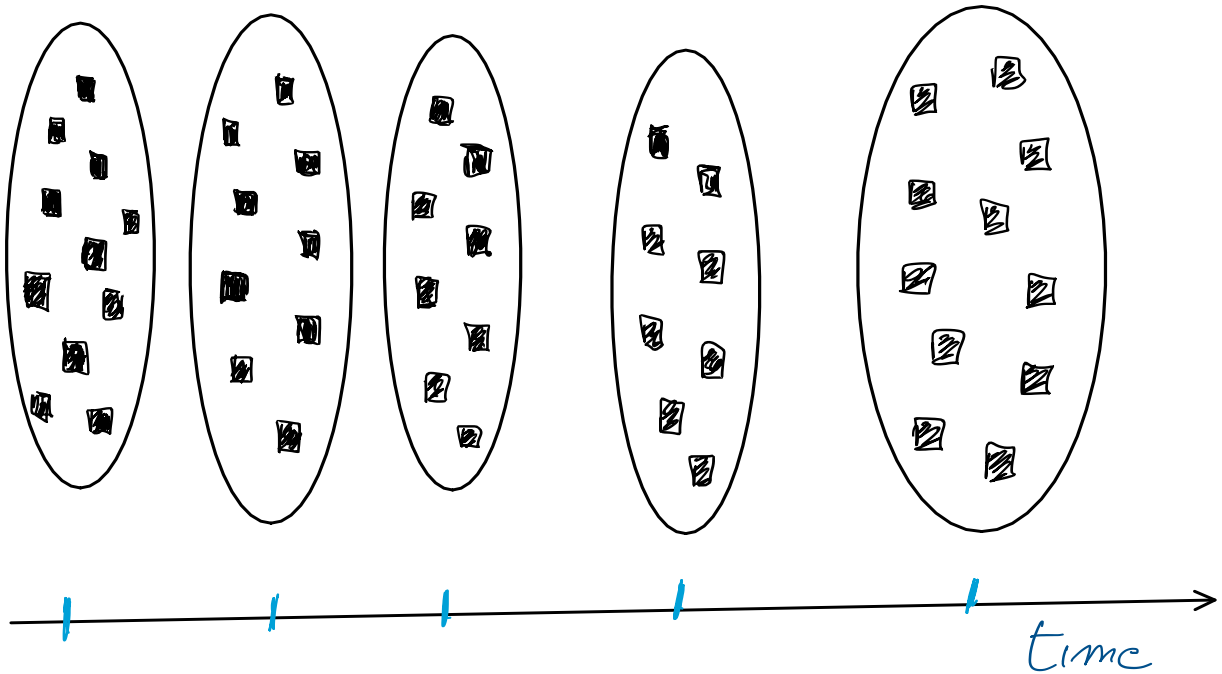
Machine Learning 4 Nutrition Science Project (ML4N)

Adele de Hoffer, Shahram Vatani, Corentin Cot, Giacomo Cacciapaglia, Maria Luisa Chiusano, Andrea Cimarelli, Francesco Conventi, Antonio Giannini, Stefan Hohenegger, Francesco Sannino

Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19.

Sci Rep **12**, 9275 (2022). https://doi.org/10.1038/s41598-022-12442-8

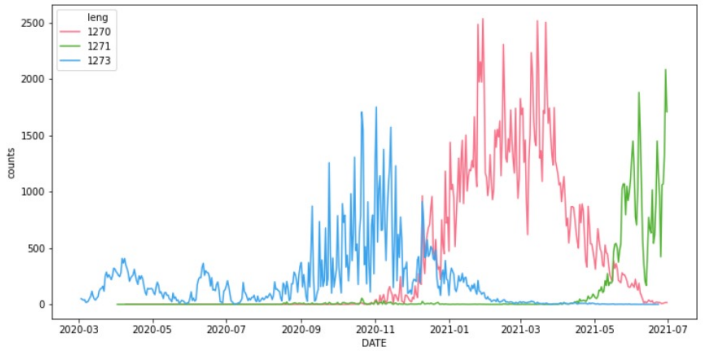
THE BASIC IDEA



Spike protein single sequence

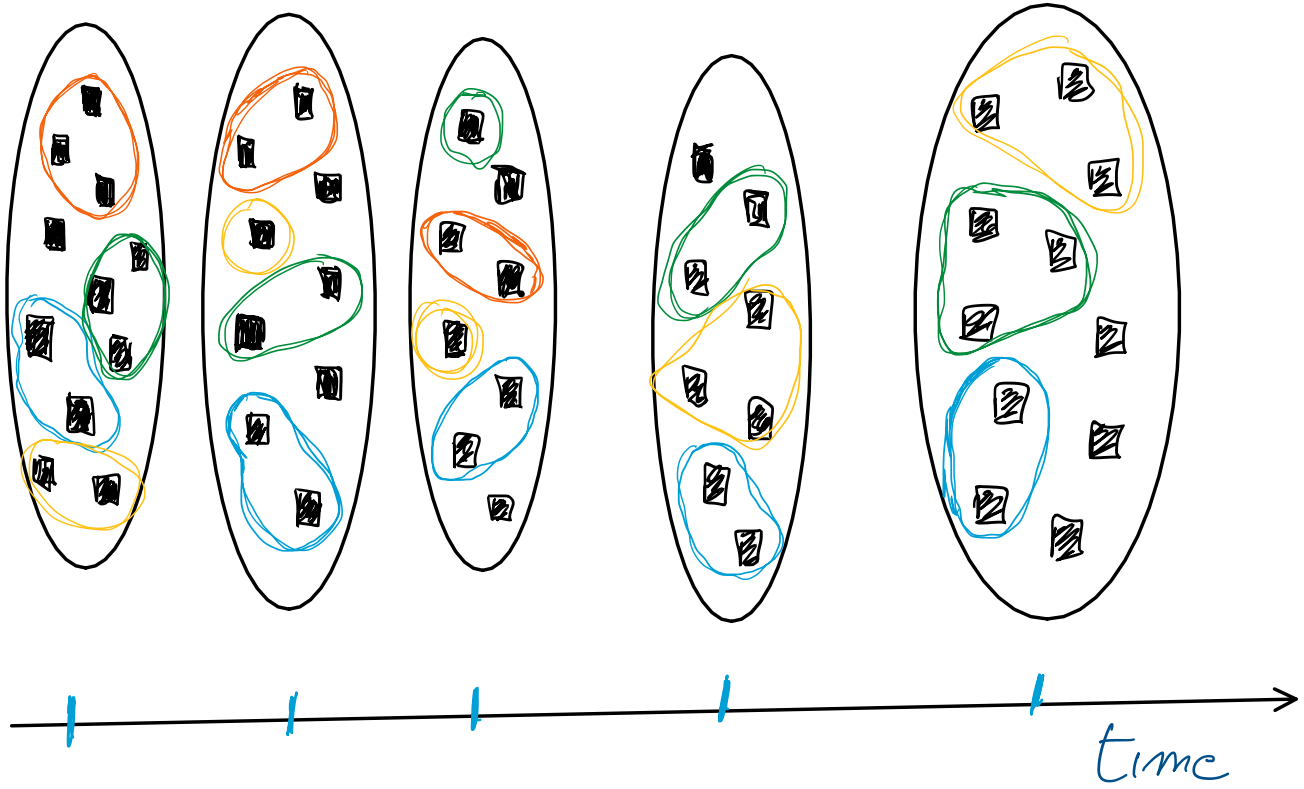
	EPI ISL	ORIGIN LAB	DATE	Sequence	Month
0	610245	Virology	2020-01-29	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	1
1	407071	Respiratory Virus Unit	2020-01-29	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	1
2	407073	Respiratory Virus Unit	2020-01-29	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	1
3	464304	Respiratory Virus Unit	2020-02-03	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	2
4	459960	Respiratory Virus Unit	2020-02-03	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	2
...	...	...	...	...	...
261792	1375051	Lighthouse Lab in Glasgow	2021-03-20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	15
261793	1374351	Lighthouse Lab in Glasgow	2021-03-20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	15
261794	1374374	Lighthouse Lab in Glasgow	2021-03-20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	15
261795	1374391	Lighthouse Lab in Glasgow	2021-03-20	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...	15
261796	1377027	Lighthouse Lab in Cambridge	2021-03-21	MFVFLVLLPLVSSQCVNLTTRTQLPPSYTNSFTRGVYYPDKVFRSS...	15

Month	Total Sequences
3	3949
4	7349
5	2211
6	5083
7	2027
8	4433
9	9503
10	19641
11	23436
12	25351
13	46712
14	39256
15	49554
16	24061
17	24337
18	30992



(*) see: Bioinformatics, 36(9), 2020, 2697–2704

THE COOK RECIPT: 1)CLUSTERING



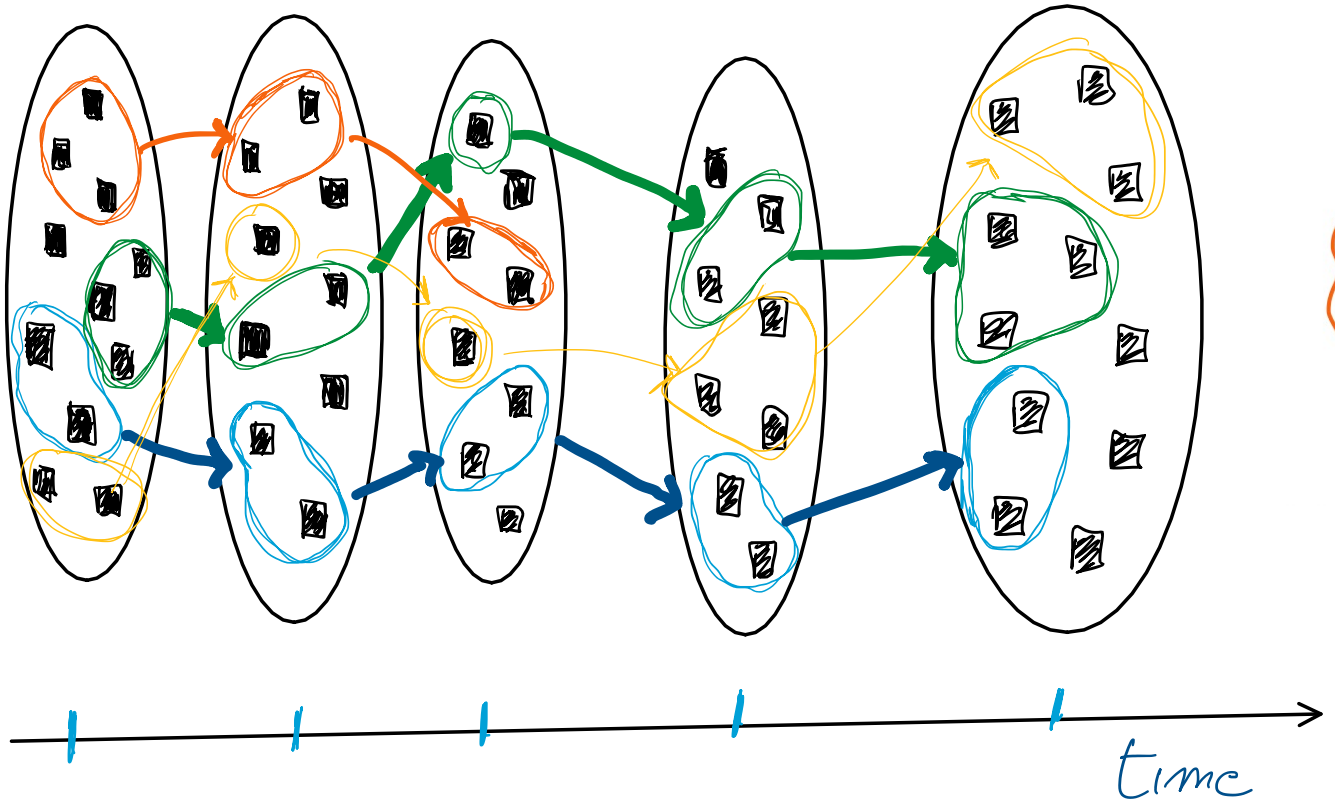
Spike protein single sequence

Cluster of sequences (at time t)

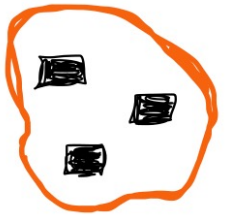
(*) see: Bioinformatics, 36(9), 2020, 2697–2704

THE COOK RECIPT:

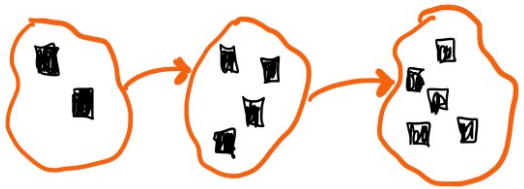
2) BUILD TIME SERIES



■ Spike protein single sequence



Cluster of sequences (at time t)

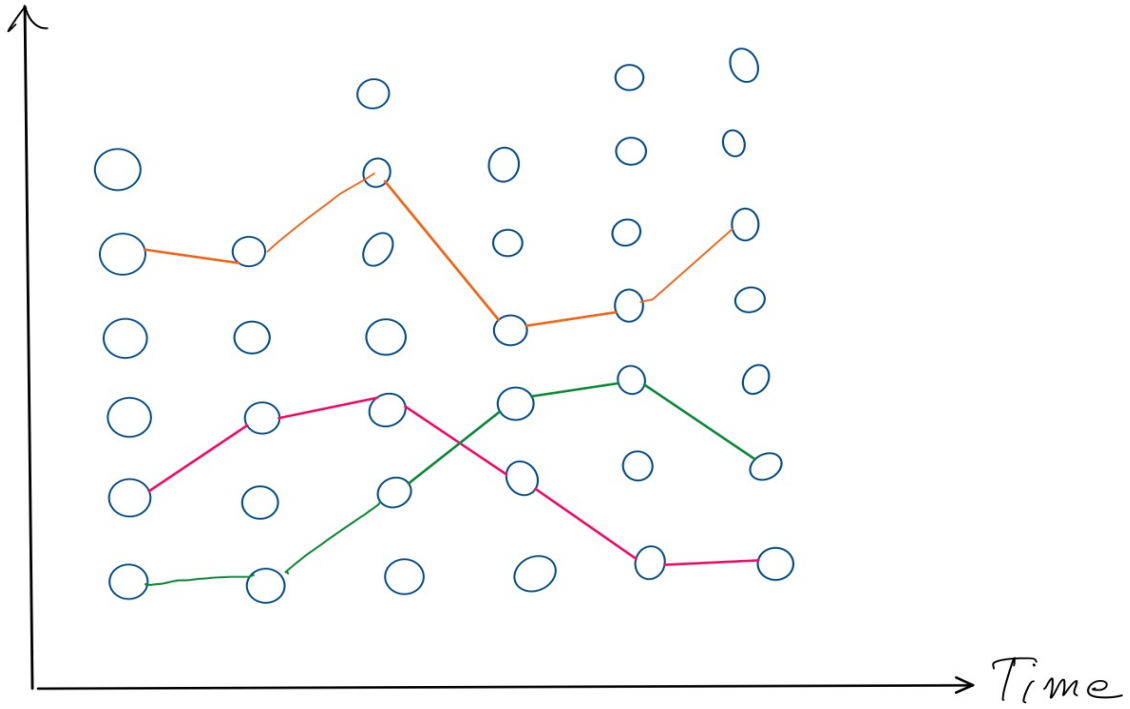


Time series of sequences

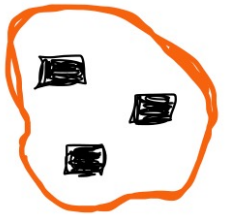
(*) see: Bioinformatics, 36(9), 2020, 2697–2704

THE COOK RECIPT:

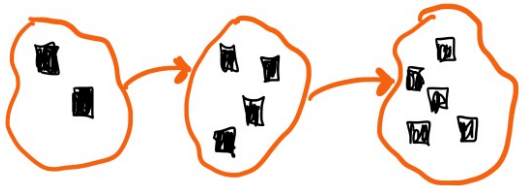
2) BUILD TIME SERIES



Spike protein single sequence



Cluster of sequences (at time t)



Time series of sequences

CLUSTERING (ON SEQUENCES)

- We need a distance defined on 1274 (or similar) chars string:

Levenshtein distance: the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other

- We need a criteria to build trees

Tree obtained aggregating elements based on thier Ward distances

- How to choose the working point?

Ward's method says that the distance between two clusters, A and B, is how much the sum of squares will increase when we merge them:

$$\Delta(A,B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \quad (2)$$

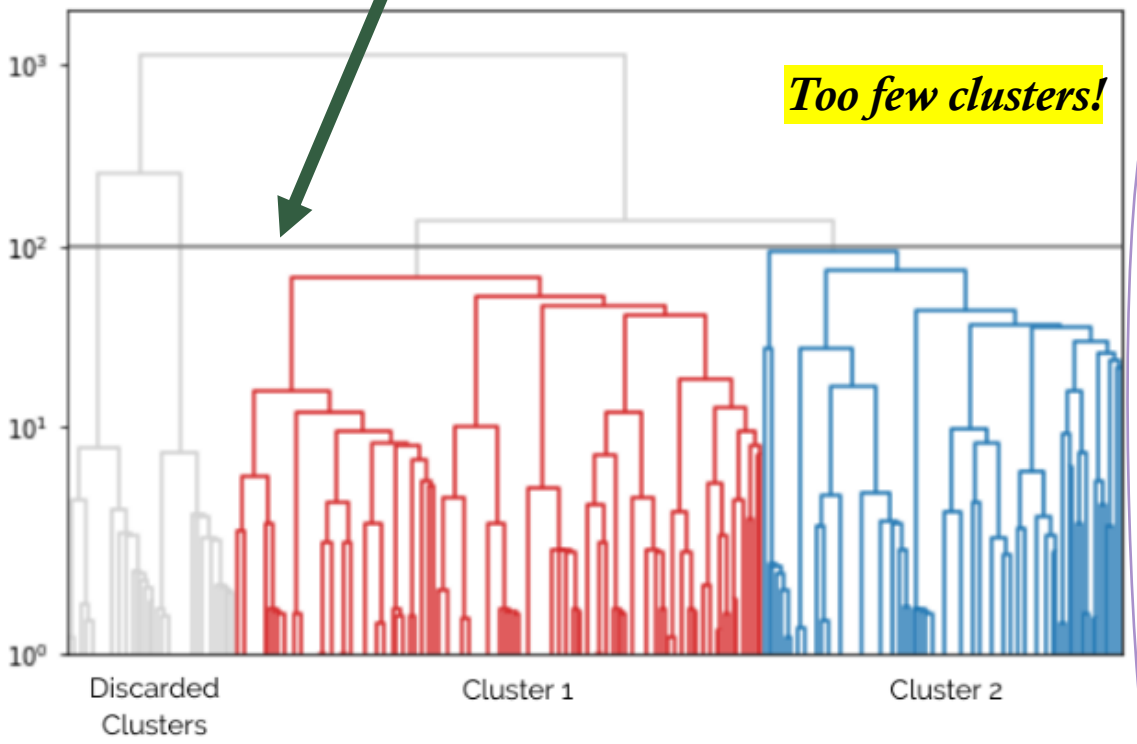
$$= \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2 \quad (3)$$

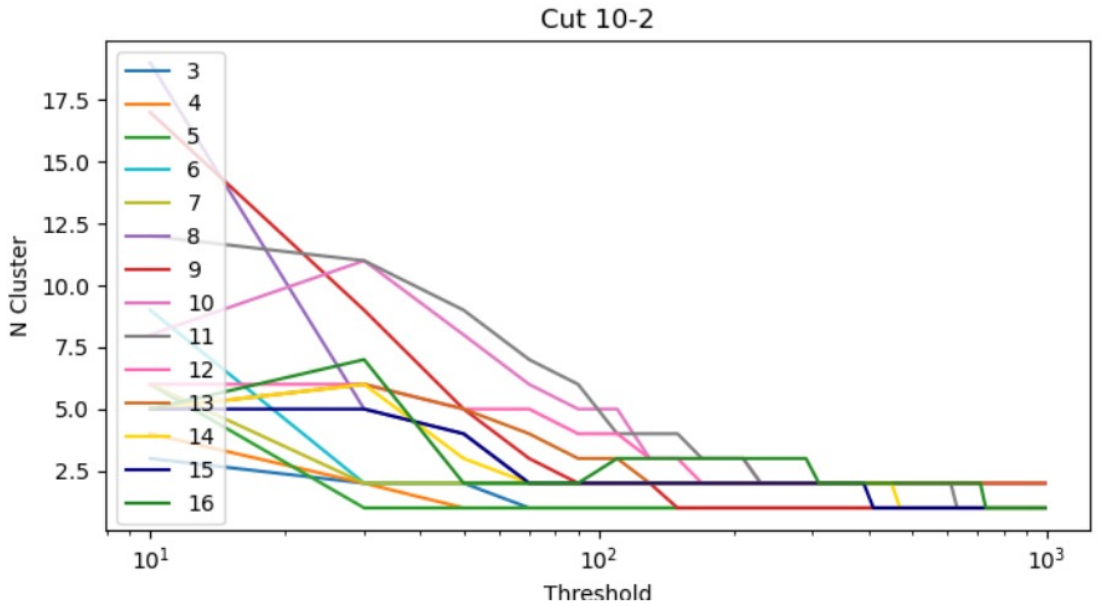
where $\vec{m}_j$ is the center of cluster j, and n_j is the number of points in it. Δ is called the **merging cost** of combining the clusters A and B.

CLUSTERING (ON SEQUENCES)

- How to choose the working point?

This line set the working point!



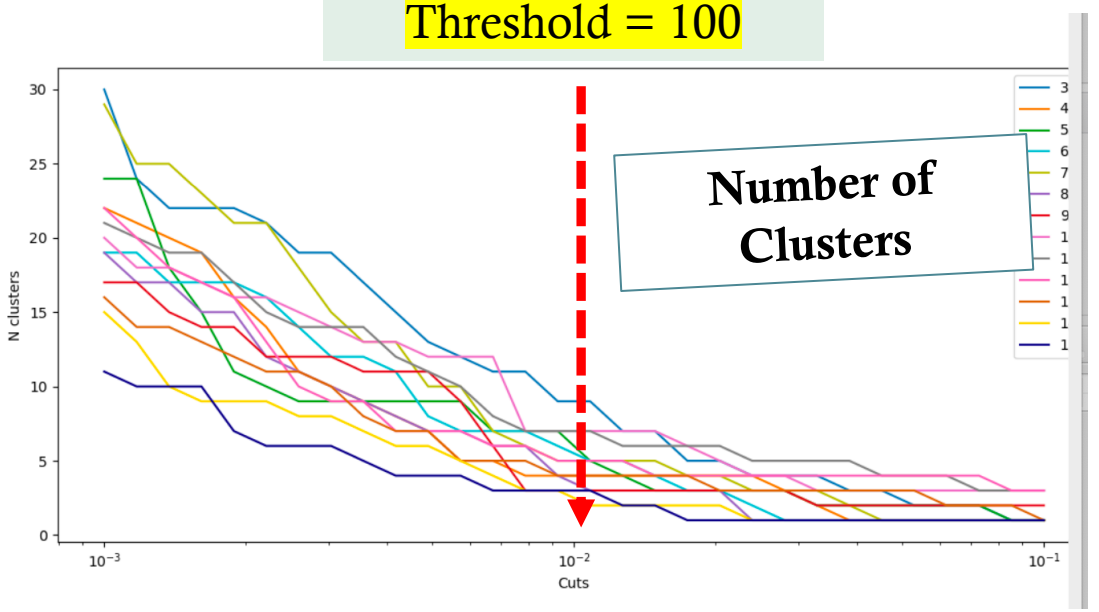


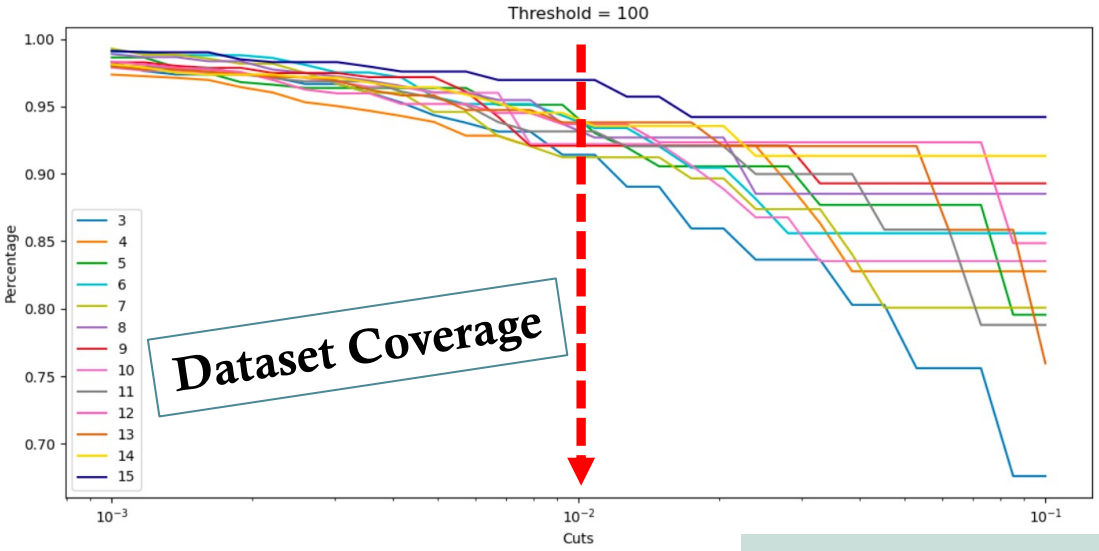
Too many clusters!

WORKING POINTS

We need to find the right **tradeoff** between the **number of custers** and the **percentage** of the dataset we are analysing, **varying the cut and the threshold**

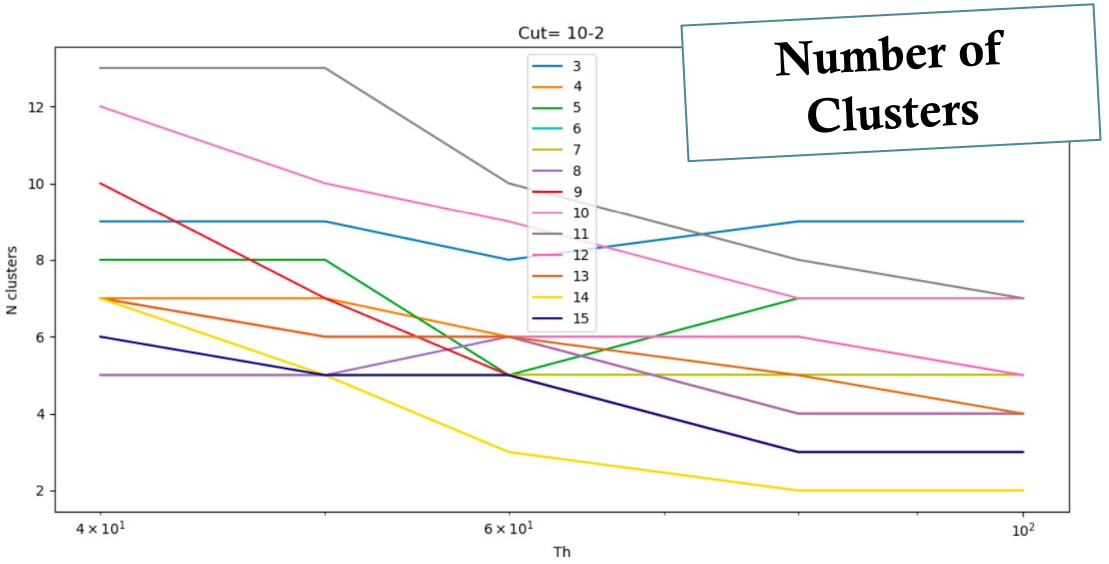
Threshold = 100

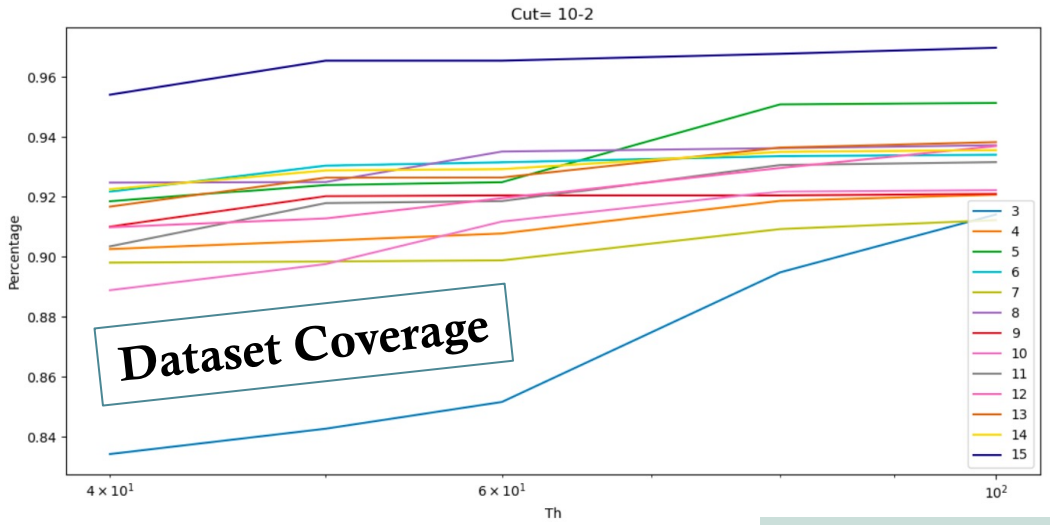




Minimum Size

WORKING POINTS





We need to find the right **tradeoff** between the **number of custers** and the **percentage** of the dataset we are analysing, **varying the cut and the threshold**

WORKING POINTS

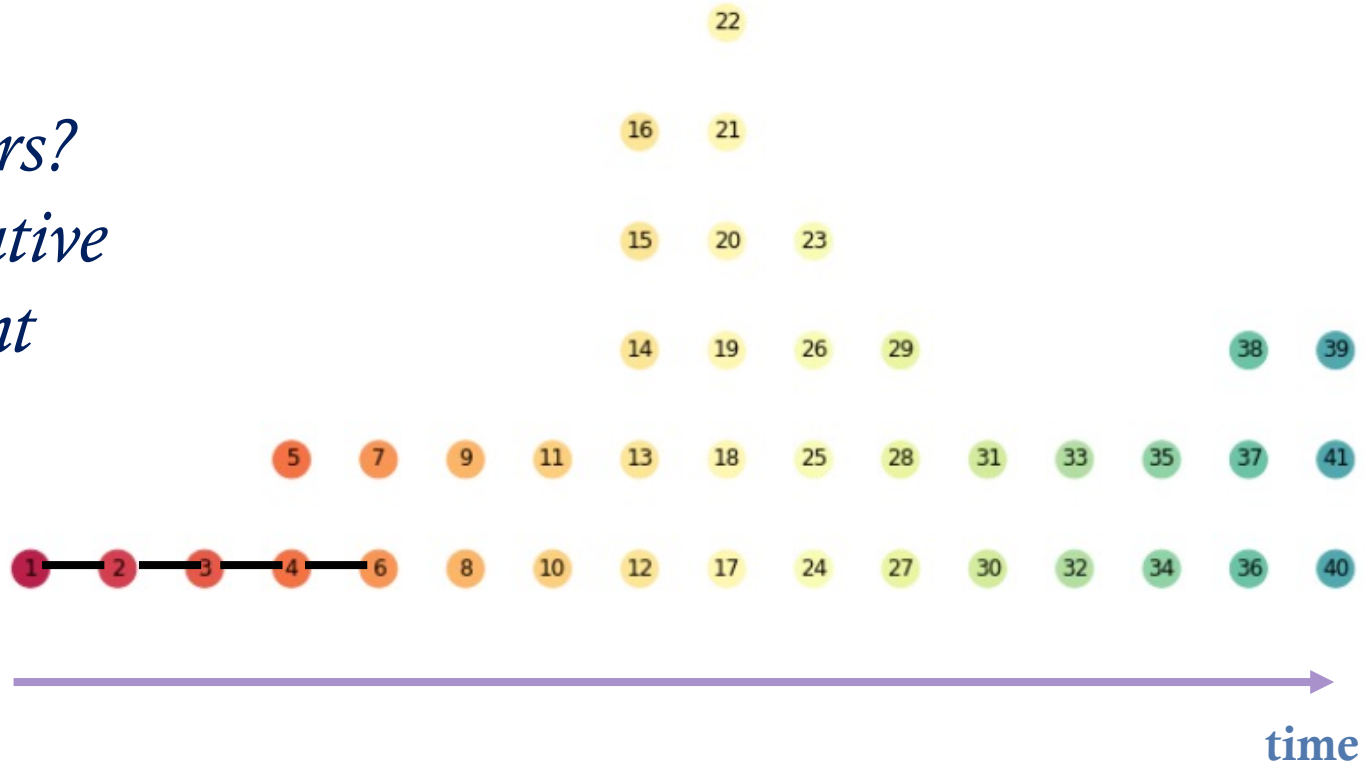
Setting:

Threshold = 100
Minimum Cluster size = 0.01

It is a good option to keep a good dataset coverage without increase too much the numbers of clusters.

HIERARCHICAL CLUSTERING RESULTS:

How to link consecutive clusters?
- Connect two or more consecutive clusters if have same dominant sequence → Chain definition



HIERARCHICAL CLUSTERING RESULTS:

How to link consecutive clusters?

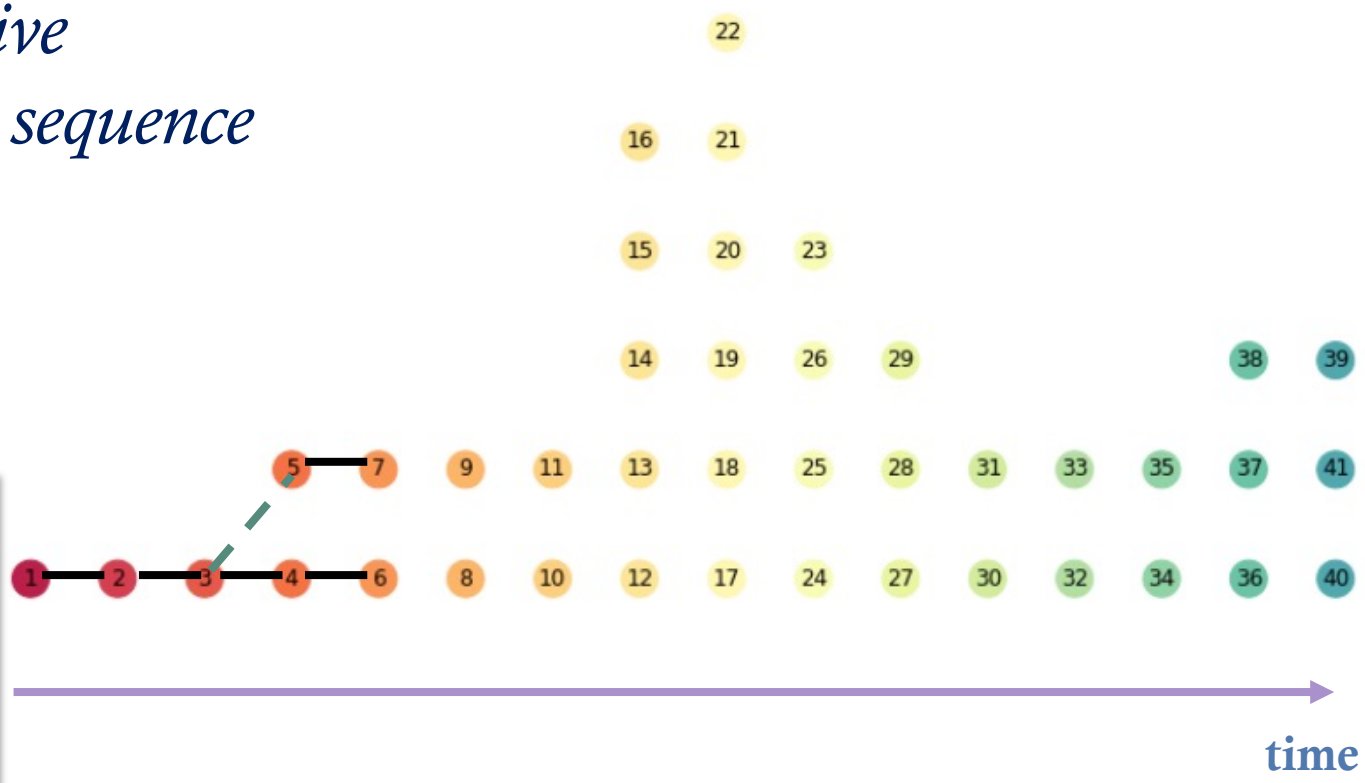
- Connect two or more consecutive clusters if have same dominant sequence

→ *Chain definition*

+ Branching algorithm!(*)

Algorithm 2 Branching algorithm for find the parenthood of cluster i

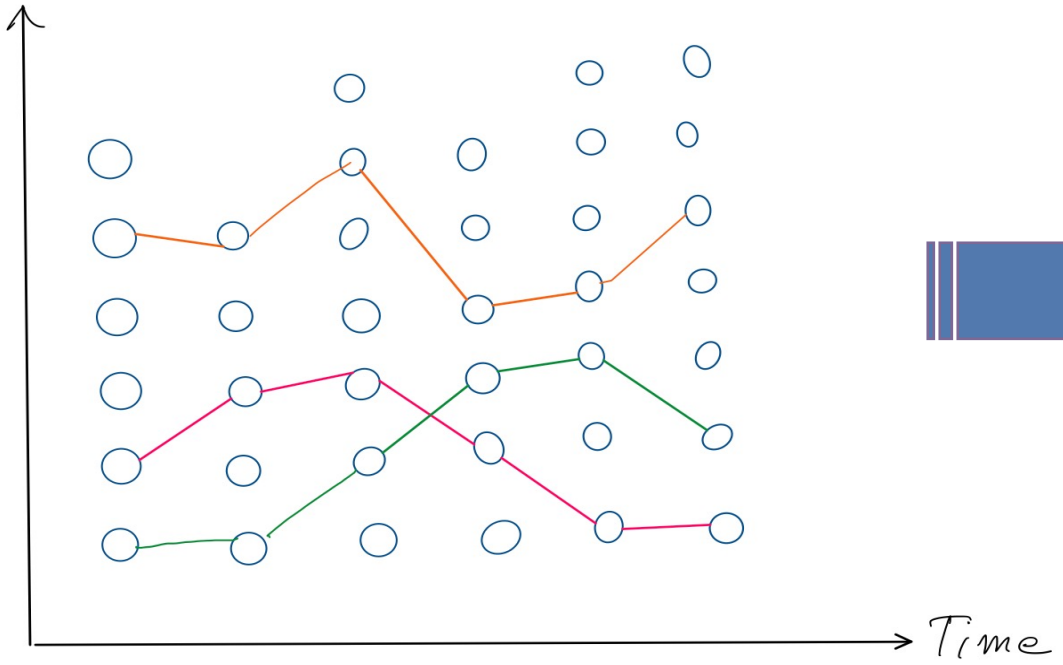
```
n= month of cluster i
s = empty array
for All the cluster k in month n-1 do
    s_k= Number of sequence in common between cluster i and cluster k
end for
if s not empty then
    Look at the maximum value of s_k
    k is the parent of i
end if
```

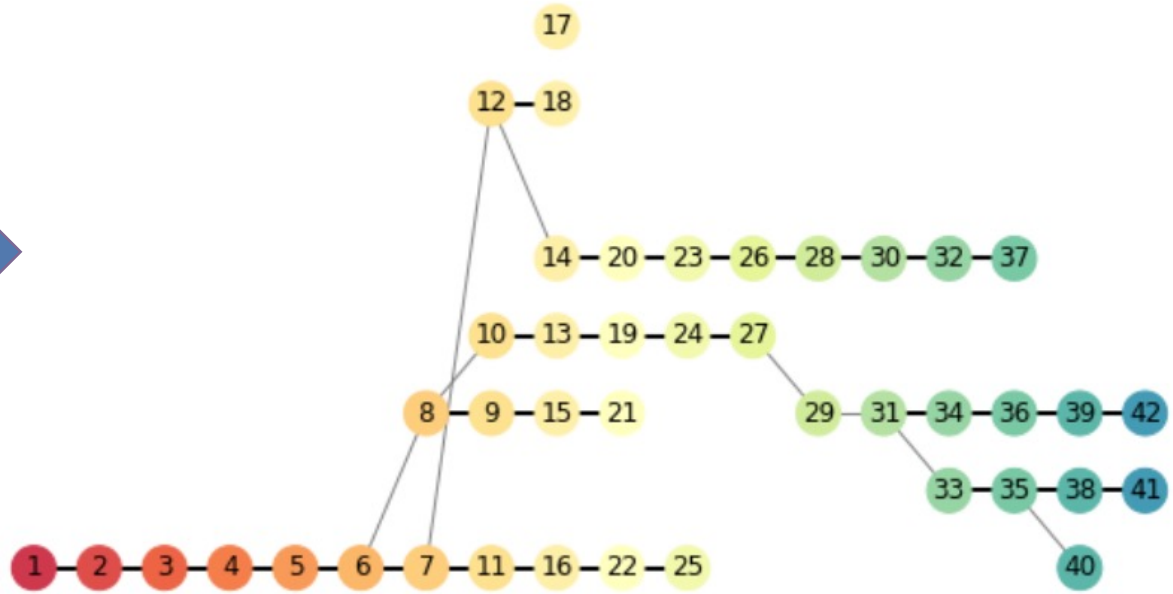


(*) (inherithed from HEP MC truth parentID)

SPIKE PROTEIN TIME-SERIES

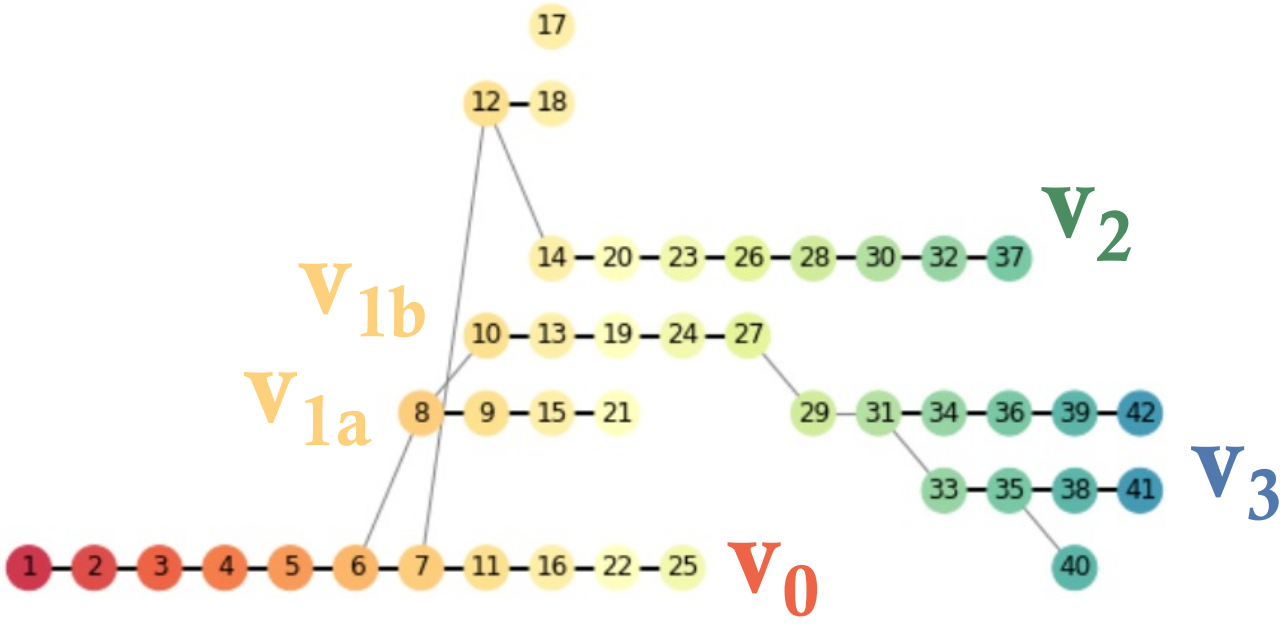
Are the dominant variants time-ordered cluster chains??

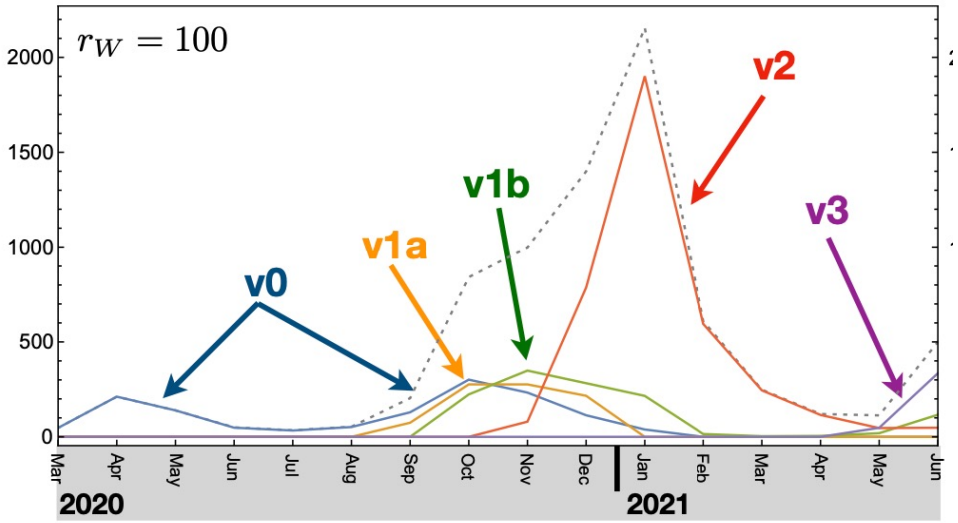


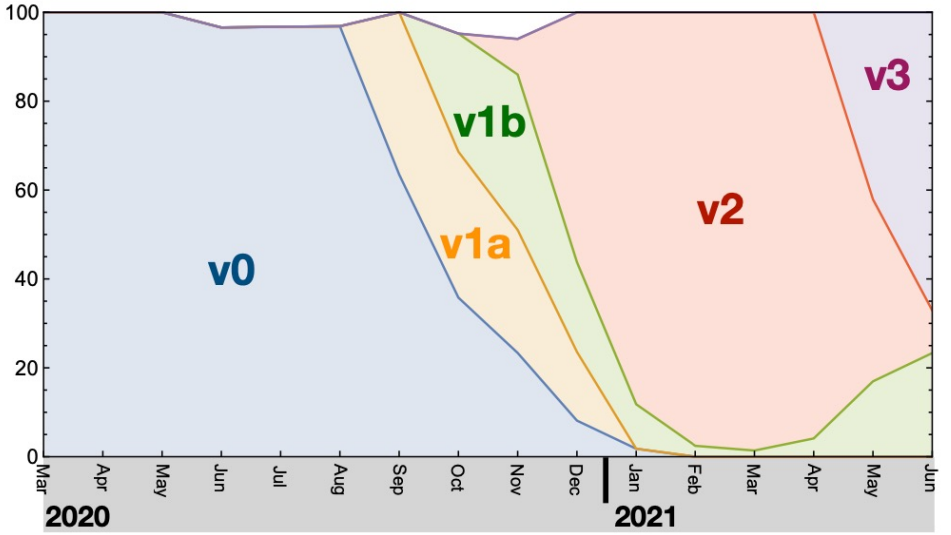


SPIKE PROTEIN TIME-SERIES

The dominant variants can be seen as time-ordered cluster chains!!

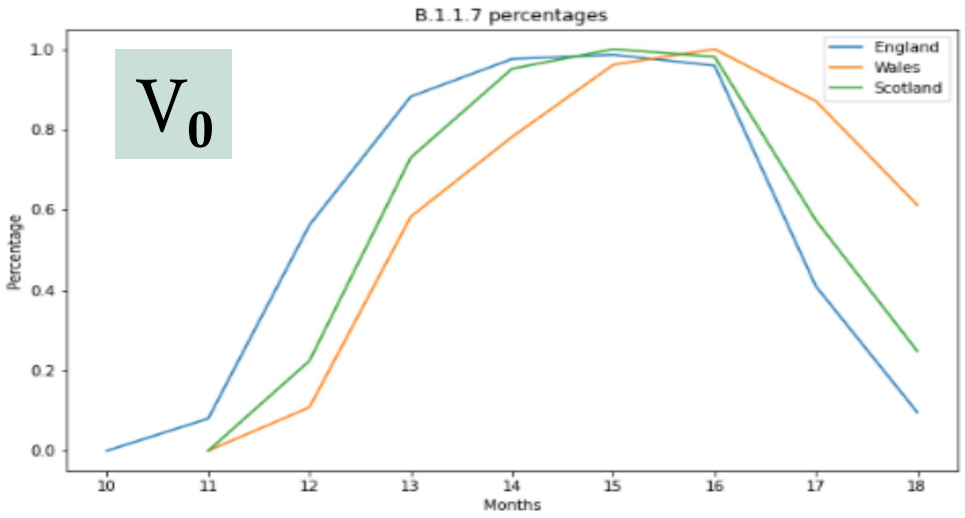


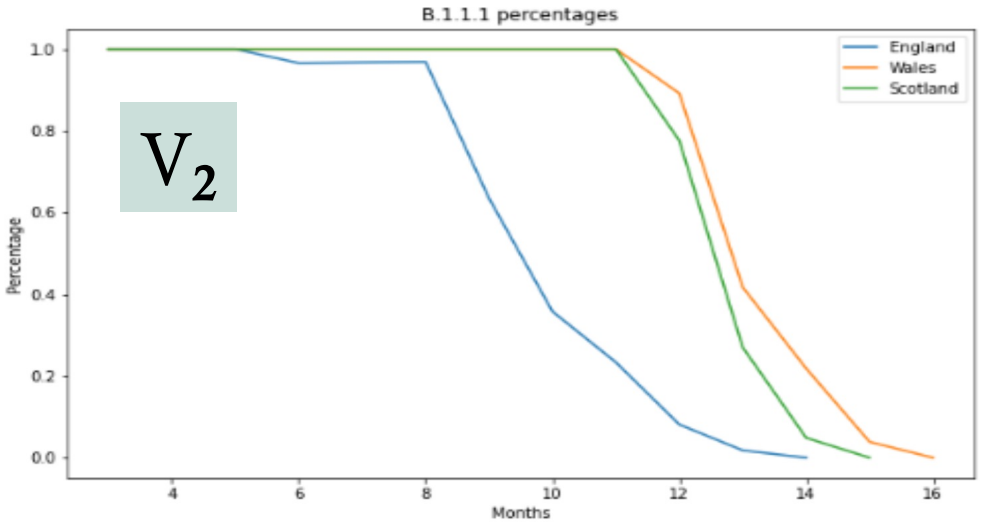




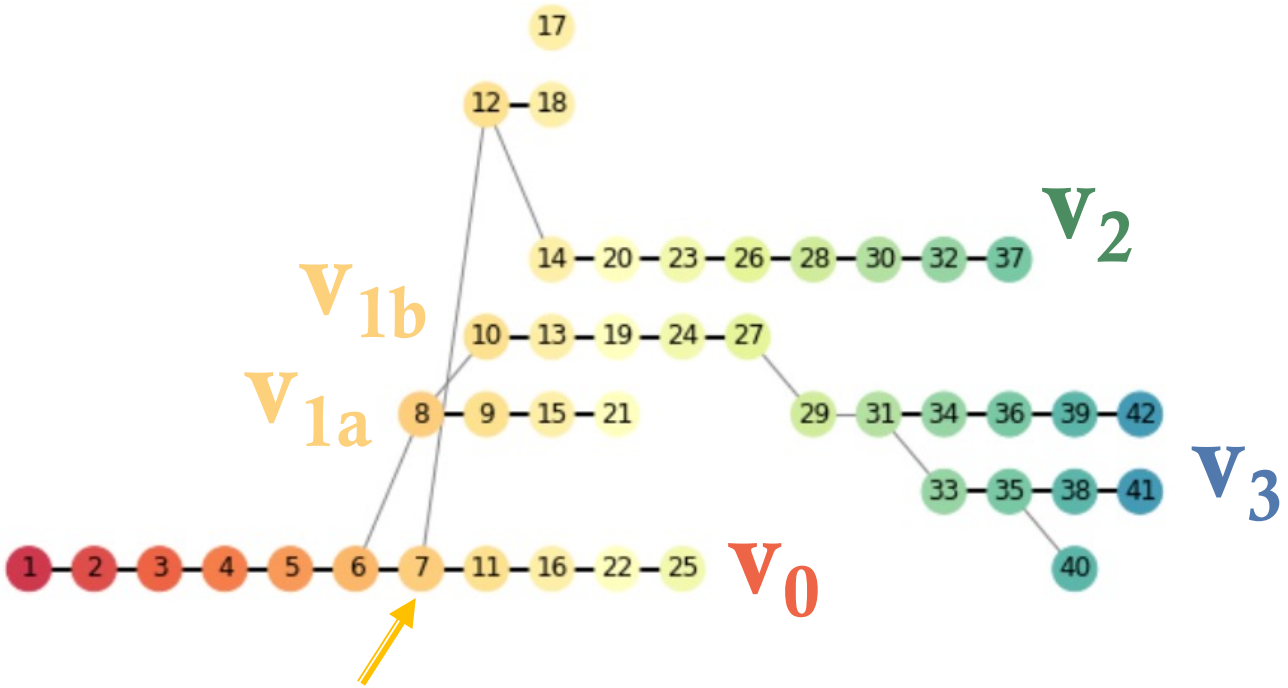
CROSS-COUNTRY VALIDATION

- Same procedure applied on Wales and Scotland dataset (*number of sequences = O(10-1)*) in blind mode!
- We kept the same working point for the clustering procedure
- Results reproduce the dominant variants spread also for Wales and Galles!





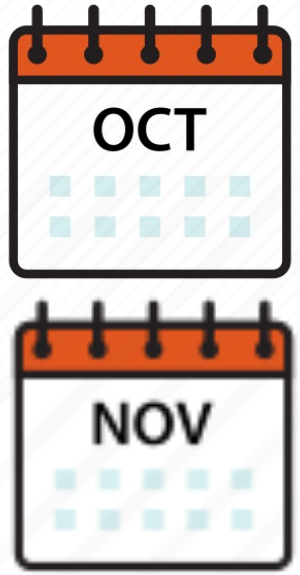
EARLY WARNING TOOL?



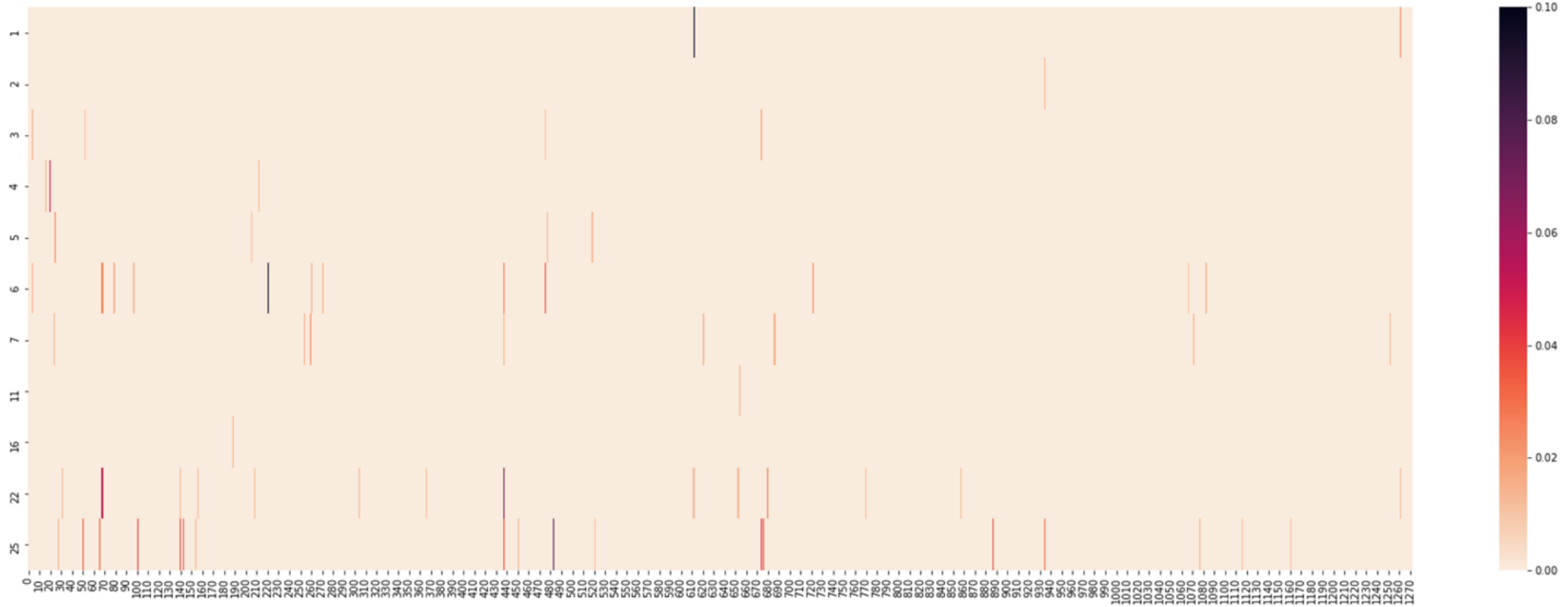
The first case of alpha variant is registered on 2020-09-20

Time resolution here is 1 month..we can try to reduce the time binning…

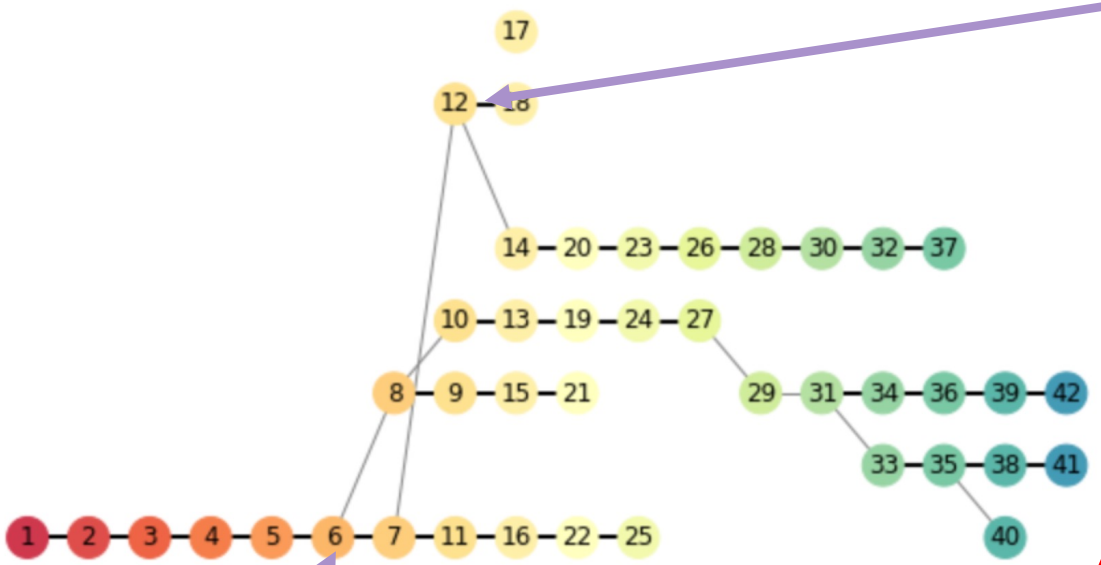
week	Total sequences	Number of alpha
38	1948	1.0
39	3394	2.0
40	2203	2.0
41	3891	4.0
42	4598	3.0
43	5921	24.0
44	4557	68.0
45	7589	225.0
46	7200	513.0
47	4669	551.0
48	2343	288.0

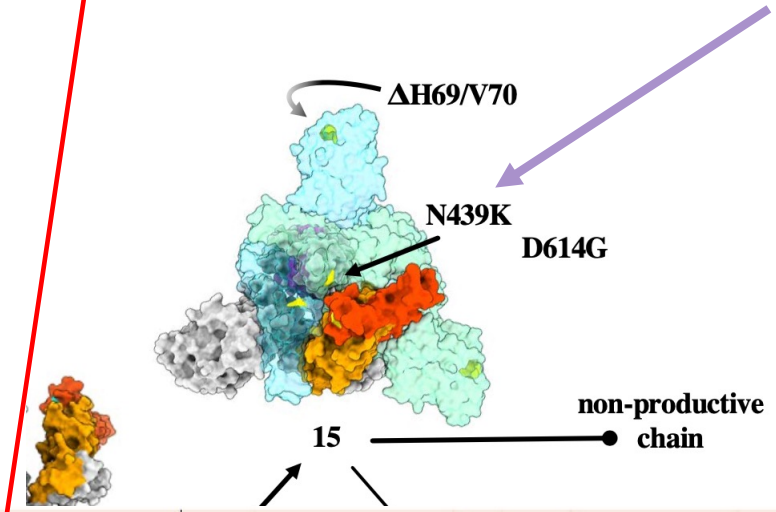


HEAT MAP FOR TIME ORDERED CLUSTER CHAIN: V0

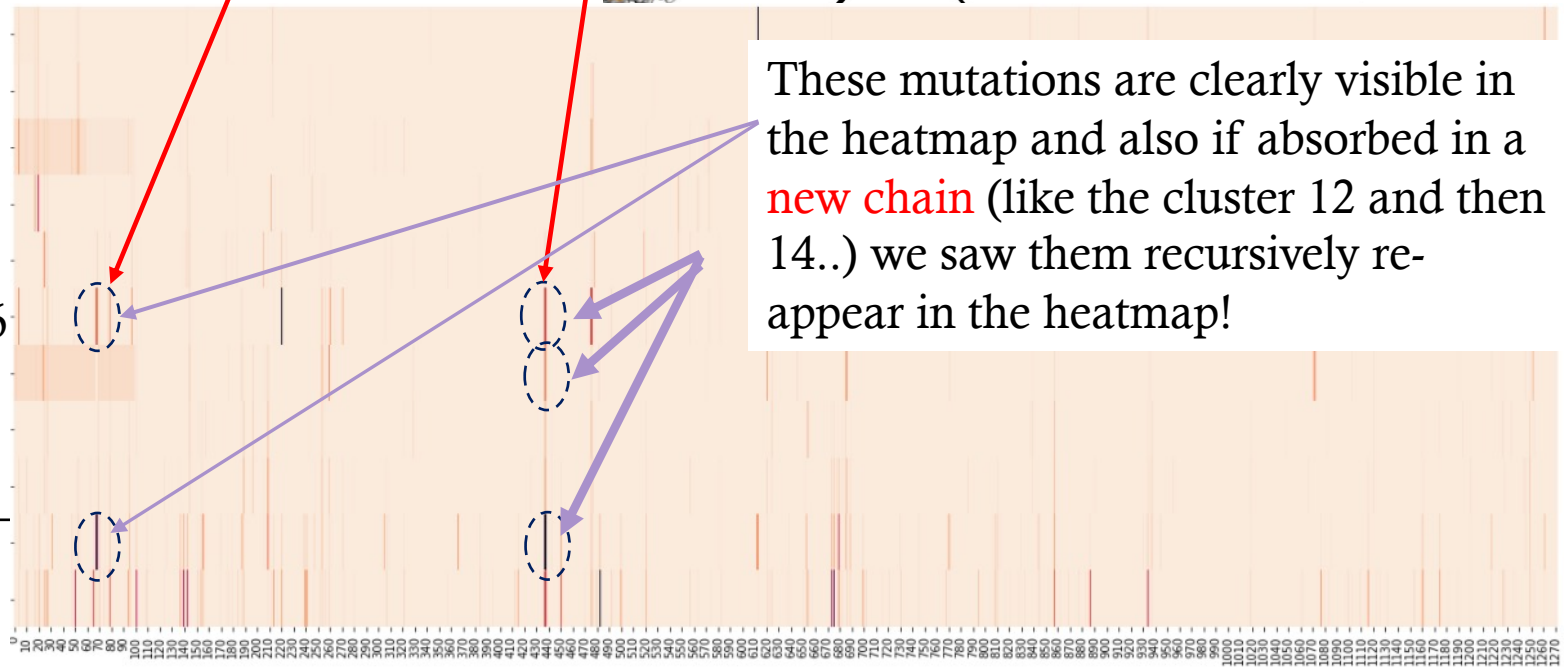


Dominant sequence in cluster 12 is:
['H69-', 'V70-', 'N439K', 'D614G'] (55% of dataset).
with 3 mutations (in red) never seen before.

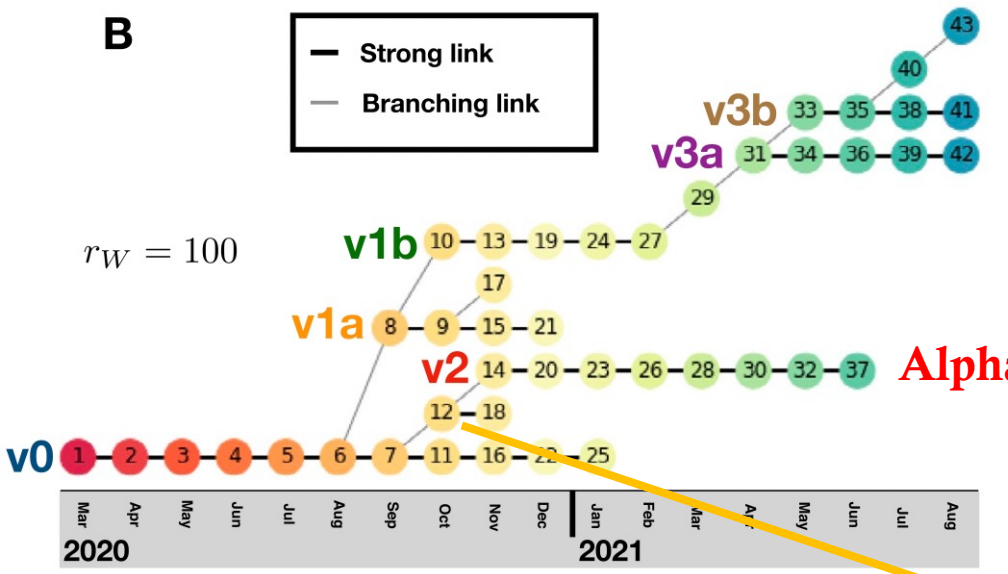




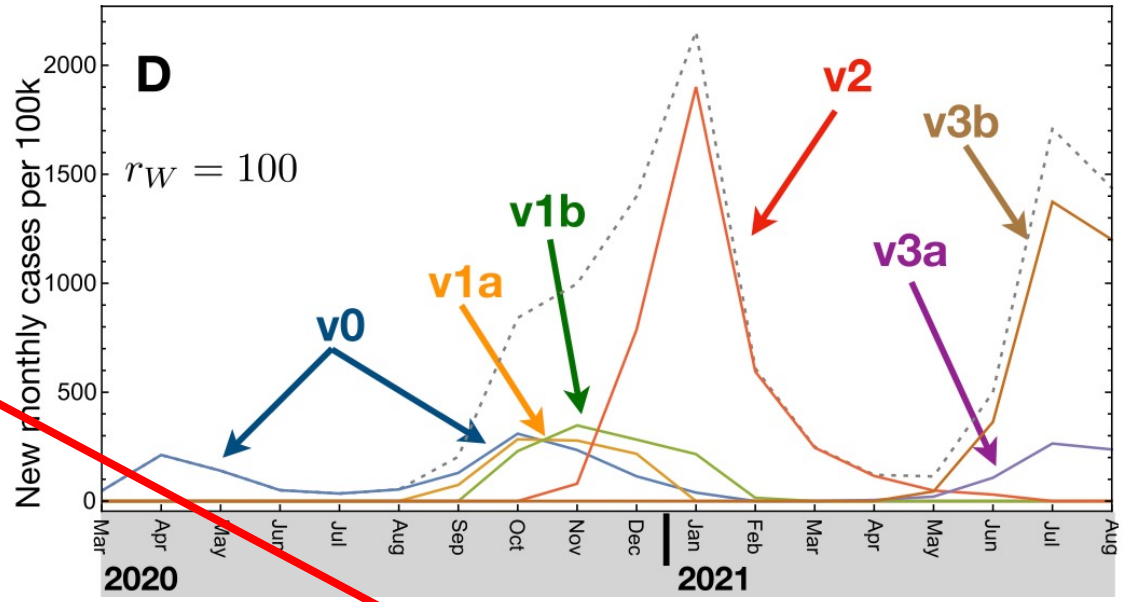
First time we found the ['H69-', 'V70-', 'N439K', 'D614G'] mutations is 2 months before in cluster 6

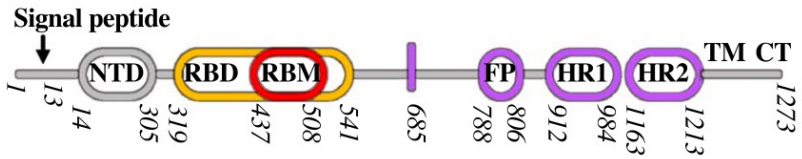


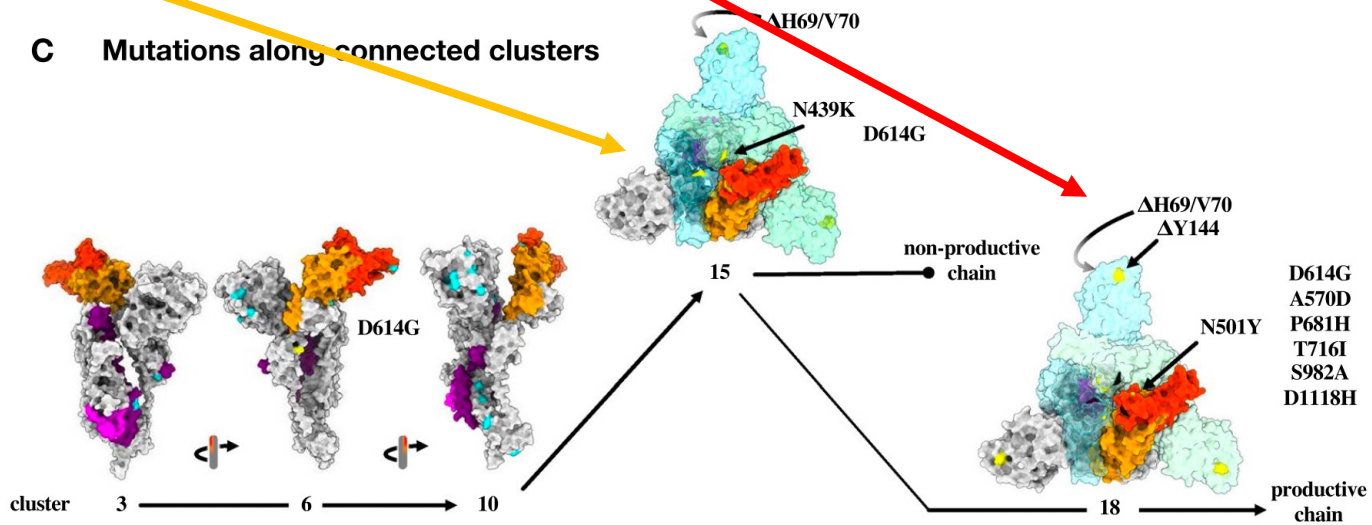
These mutations are clearly visible in the heatmap and also if absorbed in a new chain (like the cluster 12 and then 14..) we saw them recursively re-appear in the heatmap!



Alpha VoC

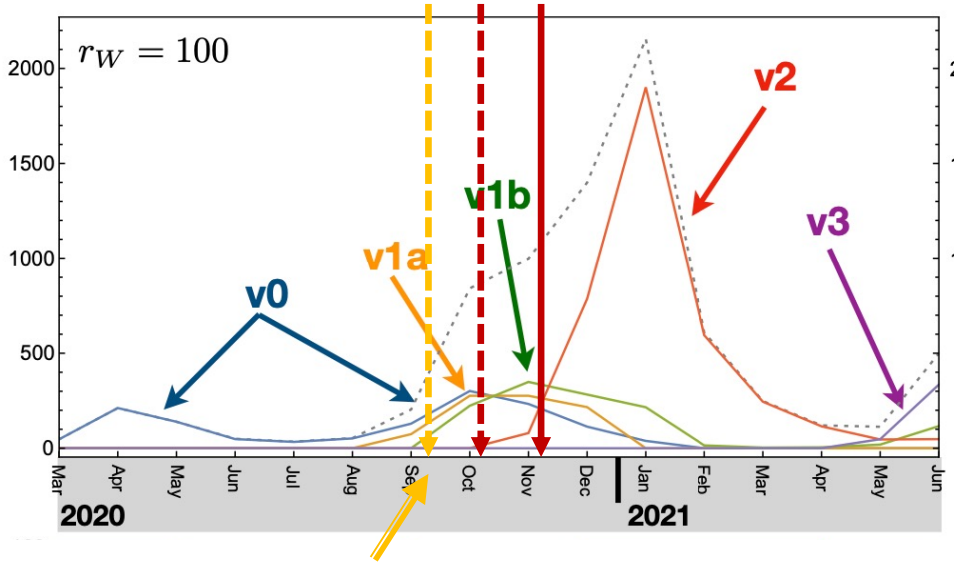




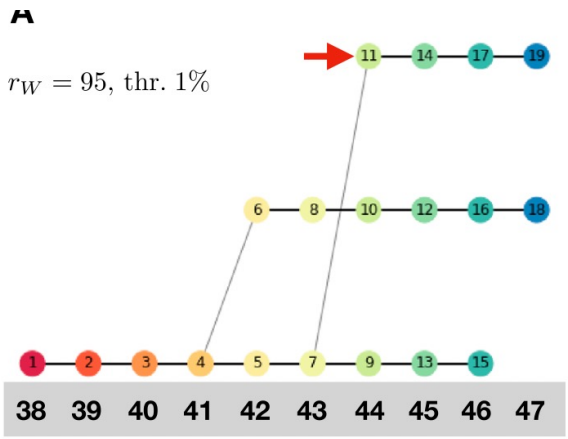


Looking at ['H69-', 'V70-', 'N439K', 'D614G'] mutations

EARLY WARNING TOOL



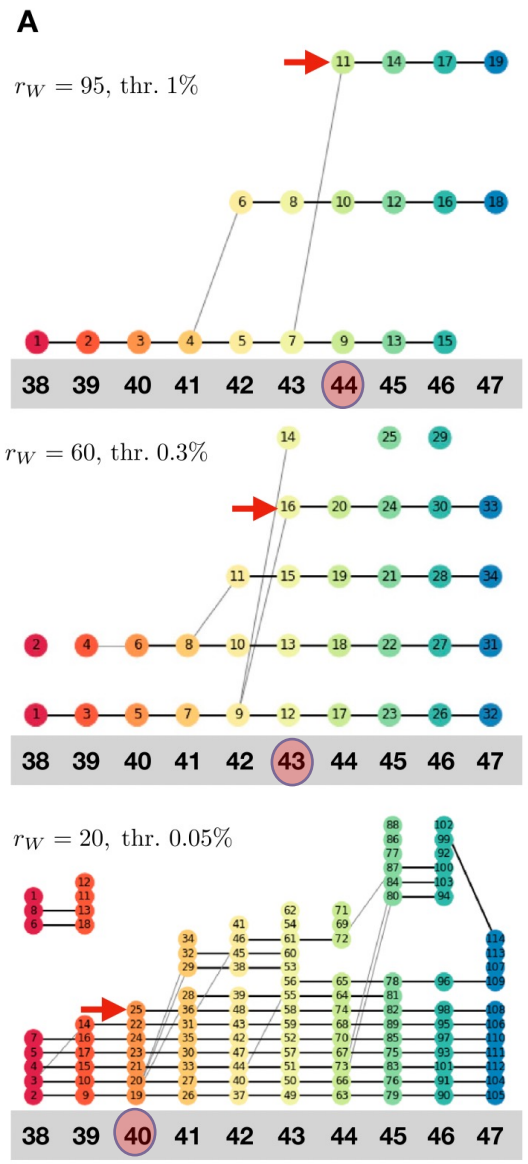
The first case of alpha variant is registered on 2020-09-20

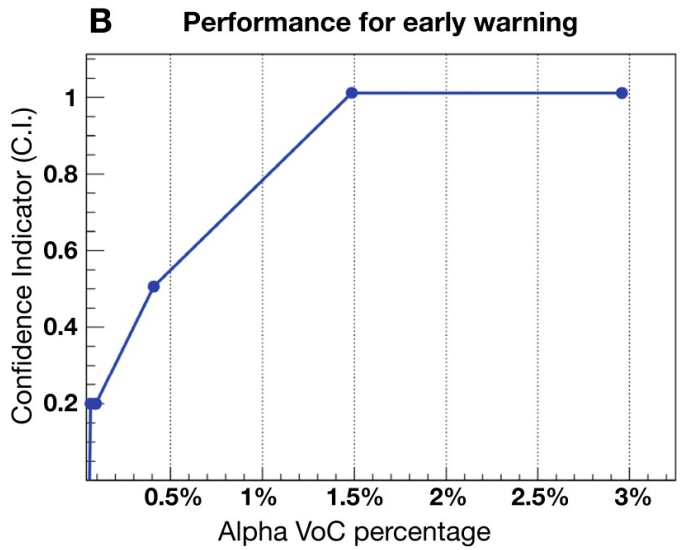


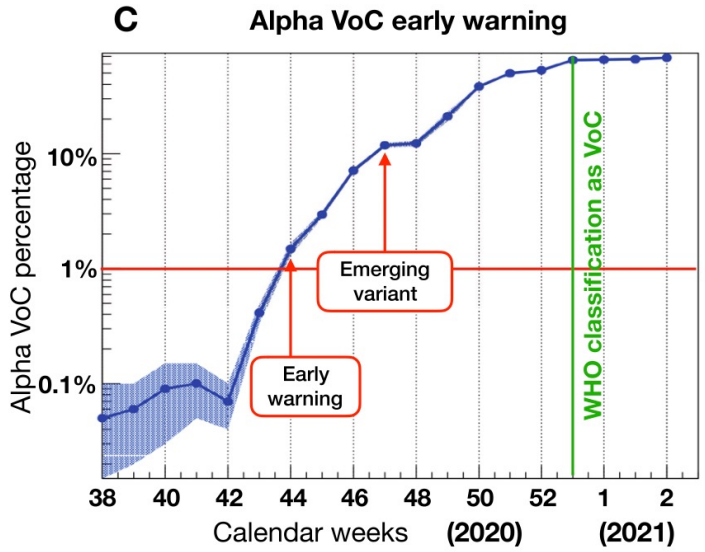
Cal. week	Date (Mon)	Total seq.	% of Alpha VoC	No. of clusters	
38	14 Sept.	1948	0.05	–	First detection
39	21 Sept.	3394	0.06	–	
40	28 Sept.	2203	0.09	5	
41	5 Oct.	3891	0.1	5	
42	12 Oct.	4598	0.07	5	
43	19 Oct.	5921	0.4	2	
44	26 Oct.	4557	1.5	1	First warning (weekly)
45	2 Nov.	7589	3	1	
46	9 Nov.	7200	7	1	
47	16 Nov.	4669	12	1	Emerging persistent variant (weekly)
48	23 Nov.	2343	12	1	
49	30Nov.	1971	21	1	First warning (monthly)
50	7 Dec.	6382	38	1	
51	14 Dec.	8059	50	1	
52	21 Dec.	4864	53	1	
53	28 Dec.	7766	65	1	WHO classification as VoC

EARLY WARNING TOOL

The first case of alpha variant is registered on 2020-09-20

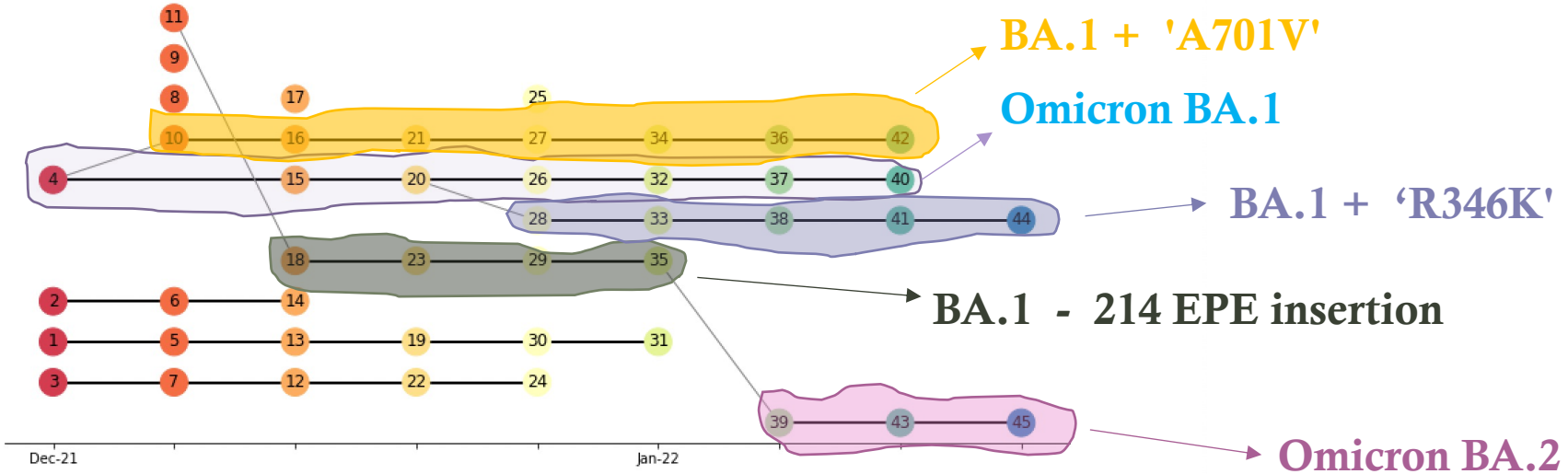


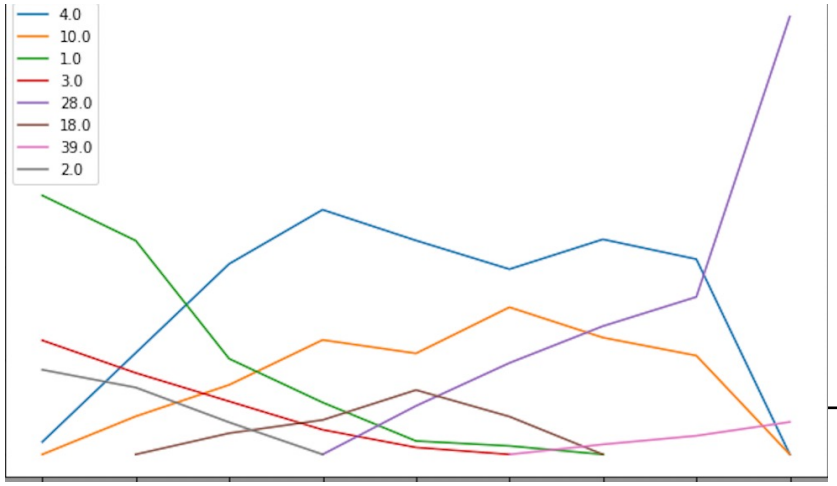




Total seq.	% of Alpha VoC	No. of clusters	
1948	0.05	–	First detection
3394	0.06	–	
2203	0.09	5	
3891	0.1	5	
4598	0.07	5	
5921	0.4	2	
4557	1.5	1	First warning (weekly)
7589	3	1	
7200	7	1	
4669	12	1	Emerging persistent variant (weekly)
2343	12	1	
1971	21	1	First warning (monthly)
6382	38	1	
8059	50	1	
4864	53	1	
7766	65	1	WHO classification as VoC

Omicron results (last update on UK data)





Chains 1, 2 and 3 are delta chains!!

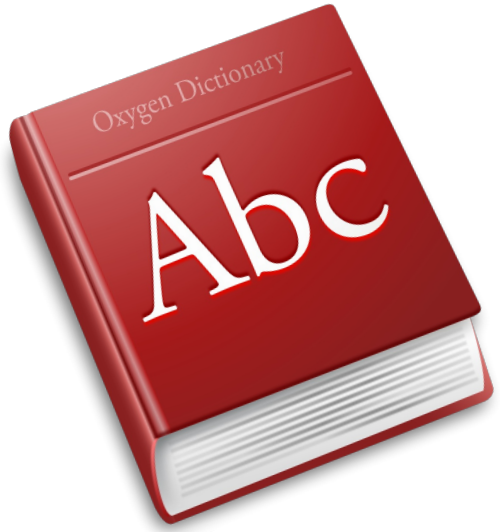
EMBEDDING AND T-SNE

Using NLP approach (→ ProtVec (*)) to represent each sequence as $\vec{a} \in R^n$ vector

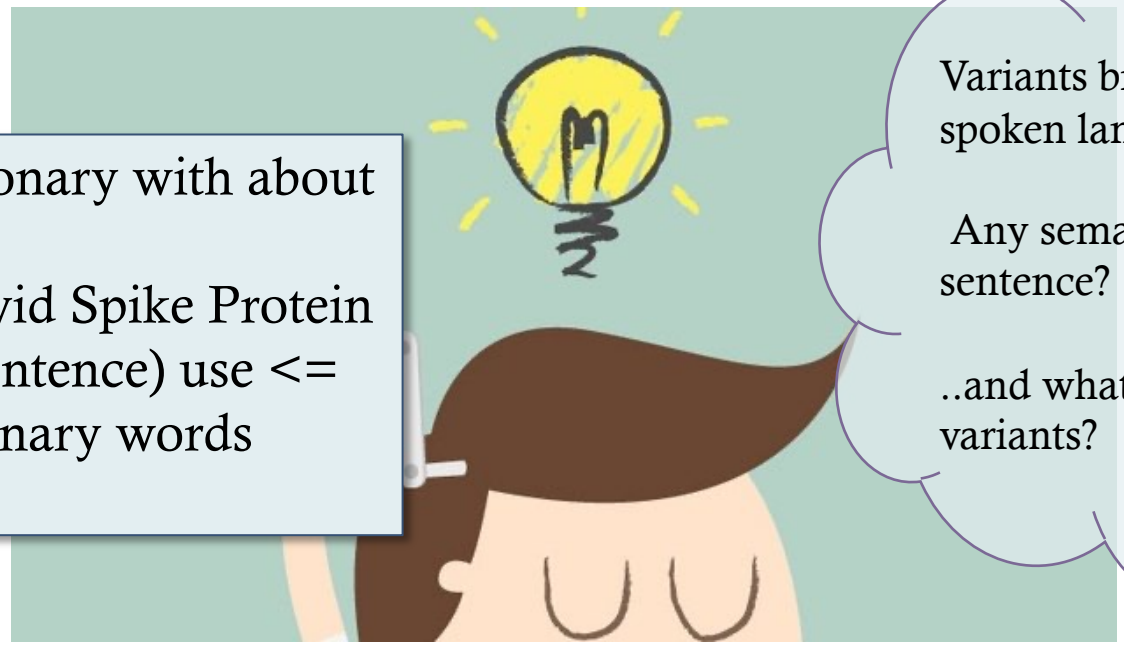
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...

Sequence as a sentence

3-grams as basic words



- Whole dictionary with about **8x10^3** words
- A single Covid Spike Protein sequence (sentence) use <= **5%** of dictionary words



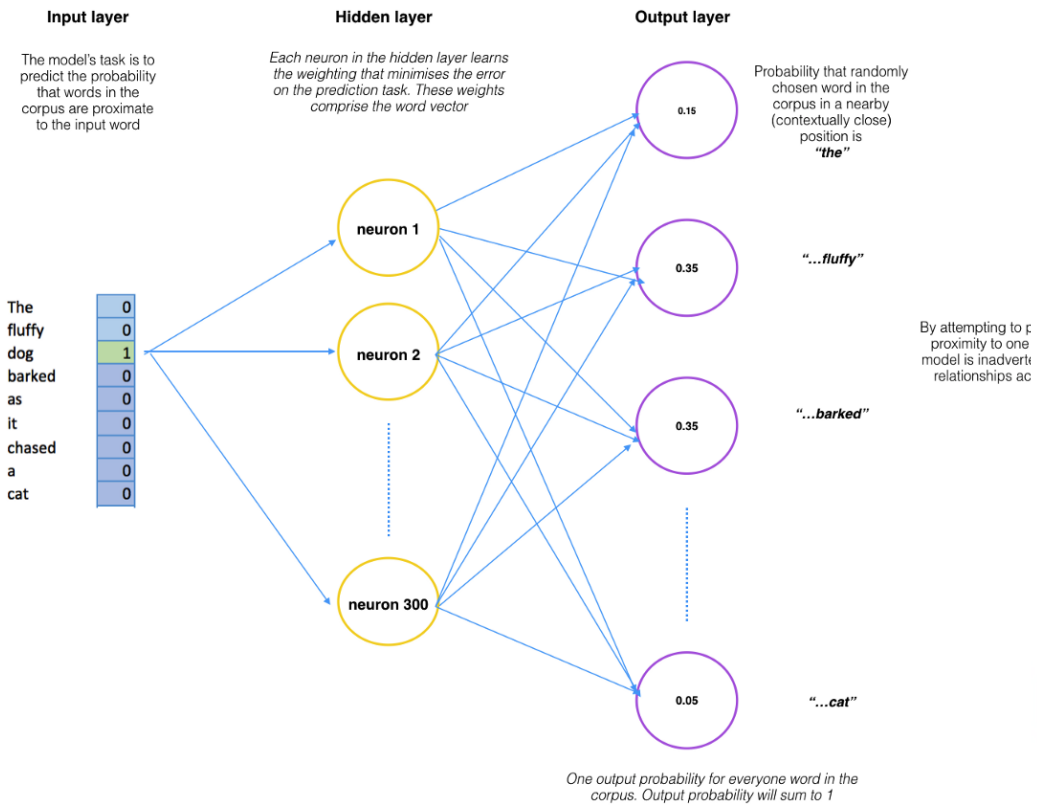
Variants bring new words to the spoken language?

Any semantic in the sequence sentence?

..and what about dominant variants?

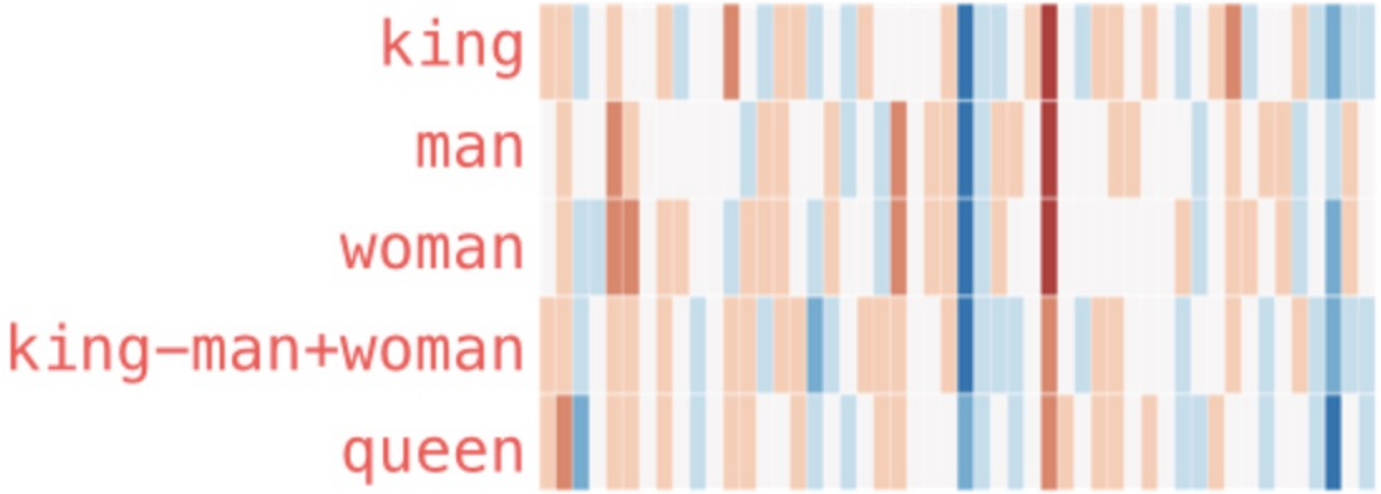
(*) PLoS ONE 10(11): e0141287, 2015

EMBEDDING AND T-SNE



MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSS...

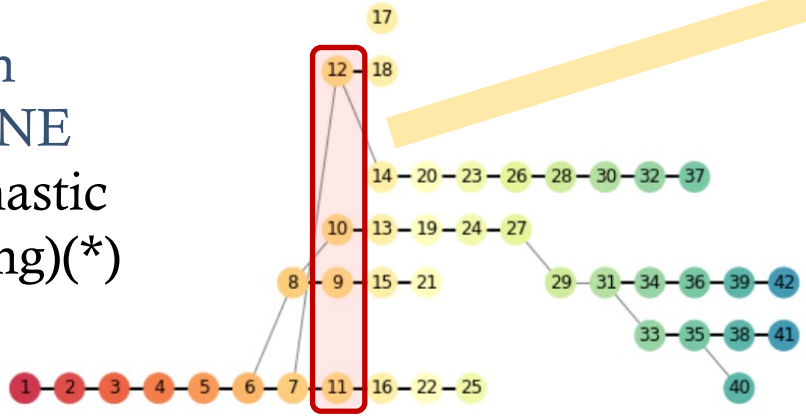
- Each 3-gram (word) is embedded in a 100 dim vector
- Weights are used to represent a 3-grams

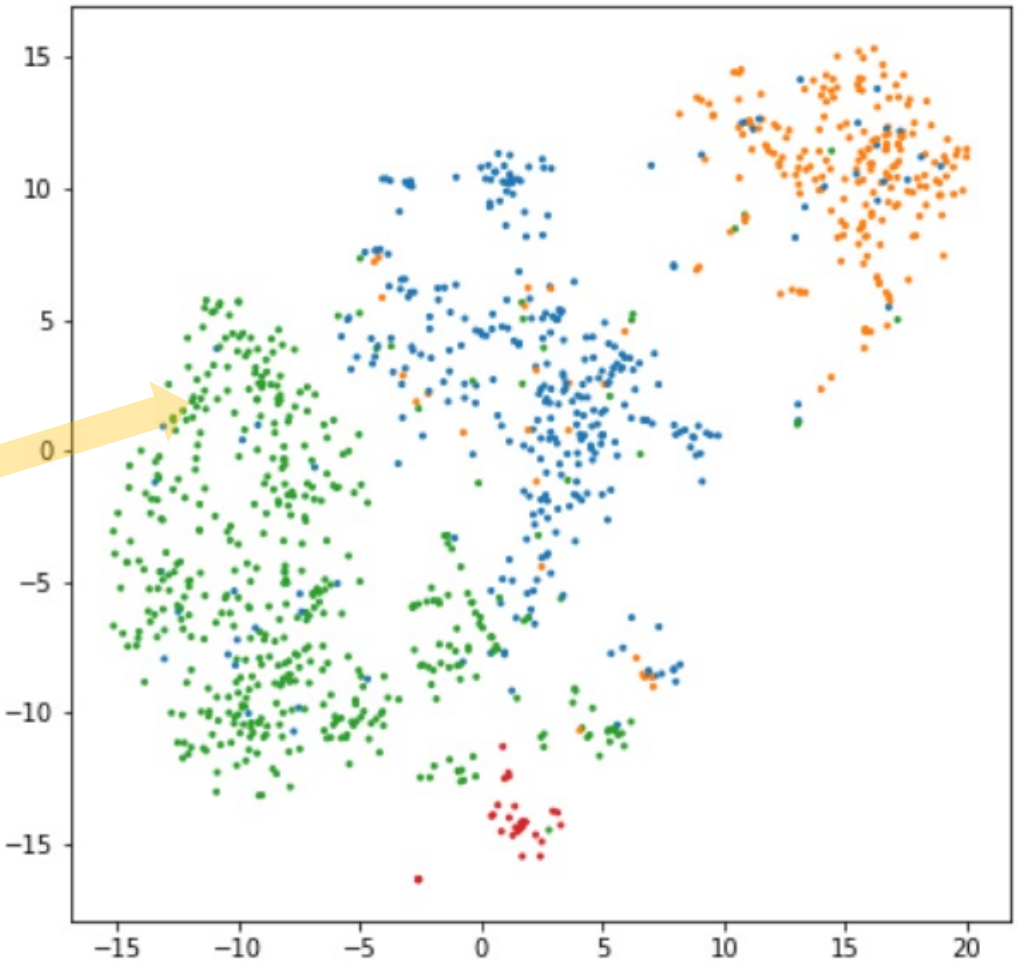


EMBEDDING VALIDATION

Is the clustering procedure based on **Levenshtein distance** equivalent to the **ProtVec** high dimensional embedding?

Dimensionality reduction and 2-dim projection using t-SNE (T-distributed Stochastic Neighbor Embedding)(*)





(*) https://lvdmaaten.github.io/tsne/

FUTURE WORK: NLP AND GRAPH NN INTERPRETATION