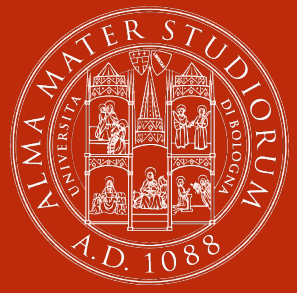


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

VEO project: a physicists' walk in the world of virology and (digital) epidemiology

Daniel Remondini

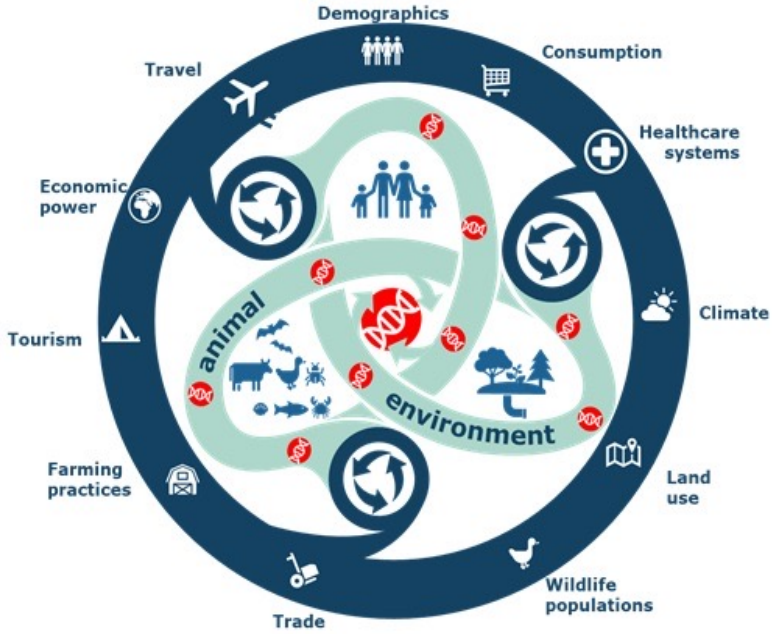
DIFA UniBO

VEO Virtual Emerging infections diseases Observatory

https://www.veo-europe.eu

Aim: develop an **interactive, virtual detection system** to monitor and analyse a wide range of information sources such as lab studies, field studies and **big data** (**genomics, geographic, social**), to be used as a possible source of information for the entities involved in controlling and limiting the spread of pathogens.

20 partners from 12 EU countries
15M euro
Jan 2020 – Dec 2025



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Epidemiological surveillance: integrated approach

Early warning on pathogens

Quantifying the impact on society

Predicting possible scenarios

Measure the effectiveness of control measures

Systematic collection, analysis and interpretation of data

- microbial/viral genomics
- human & wild animal mobility
- opinions on online social networks & media news

Digital health:
social network (Twitter) data

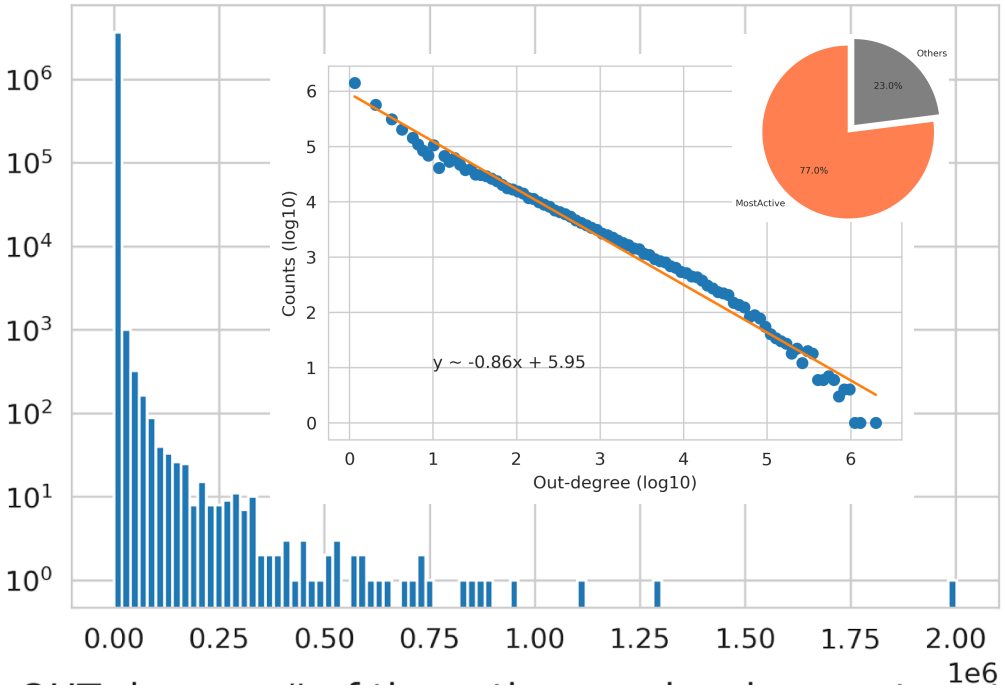
Twitter network analysis

A VEO partner (Prof. M Salathè EPFL) was collecting Tweets of COVID-related keywords (Jan-May 2020: 270M tweets ENG language) to characterize the perception and the discussions around the theme

Our proposal: apply a **network approach** to these data

Result: directed weighted network with users as nodes (22.5M) retweets as links (176M)

Hierarchical structure: 0.1% top users have >77% retweets – ROLE OF **INFLUENCERS**

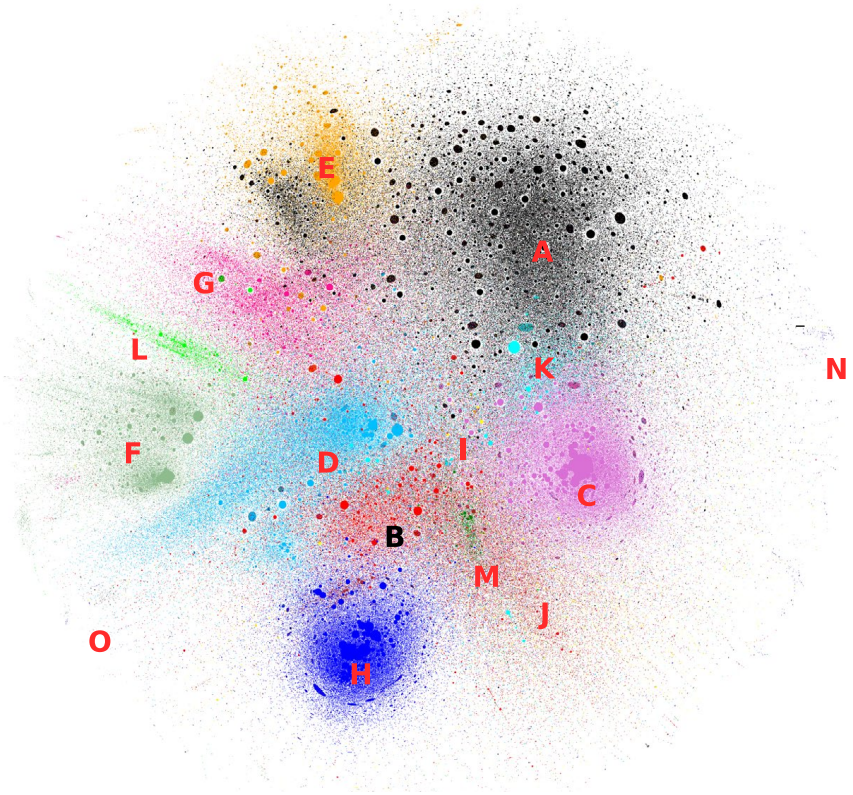


[Durazzi, ..., Remondini Sci Rep 2021]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

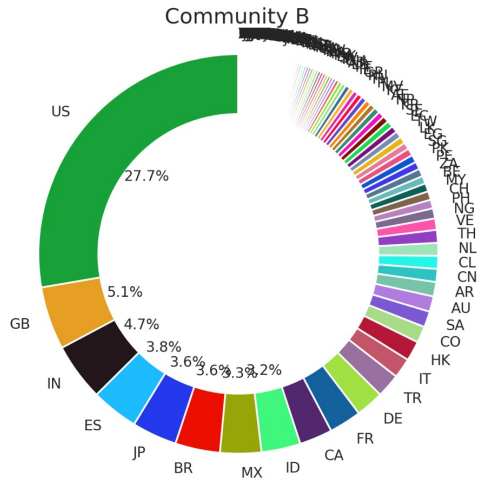
Community structure

15 large communities (>100k users each, 98% network size)
user info to identify country of origin



S	C	Name	First category	Size
		I	Sports	2.1%
		J	Science	2.0%
		D	Science	9.4%
		M	Science	0.8%
		E	Arts & Entertainment	7.2%
		N	Adult content	0.6%
		A	Arts & Entertainment	33.3%
		O	Business	0.6%
		B	Science	10.6%
		G	Science	6.4%
		F	Science	6.9%
		L	Science	1.1%
		H	Political Supporter	5.4%
		C	Political Supporter	10.0%
		K	Arts & Entertainment	1.9%

Super-community
National elite
Other
International sci-health
Political

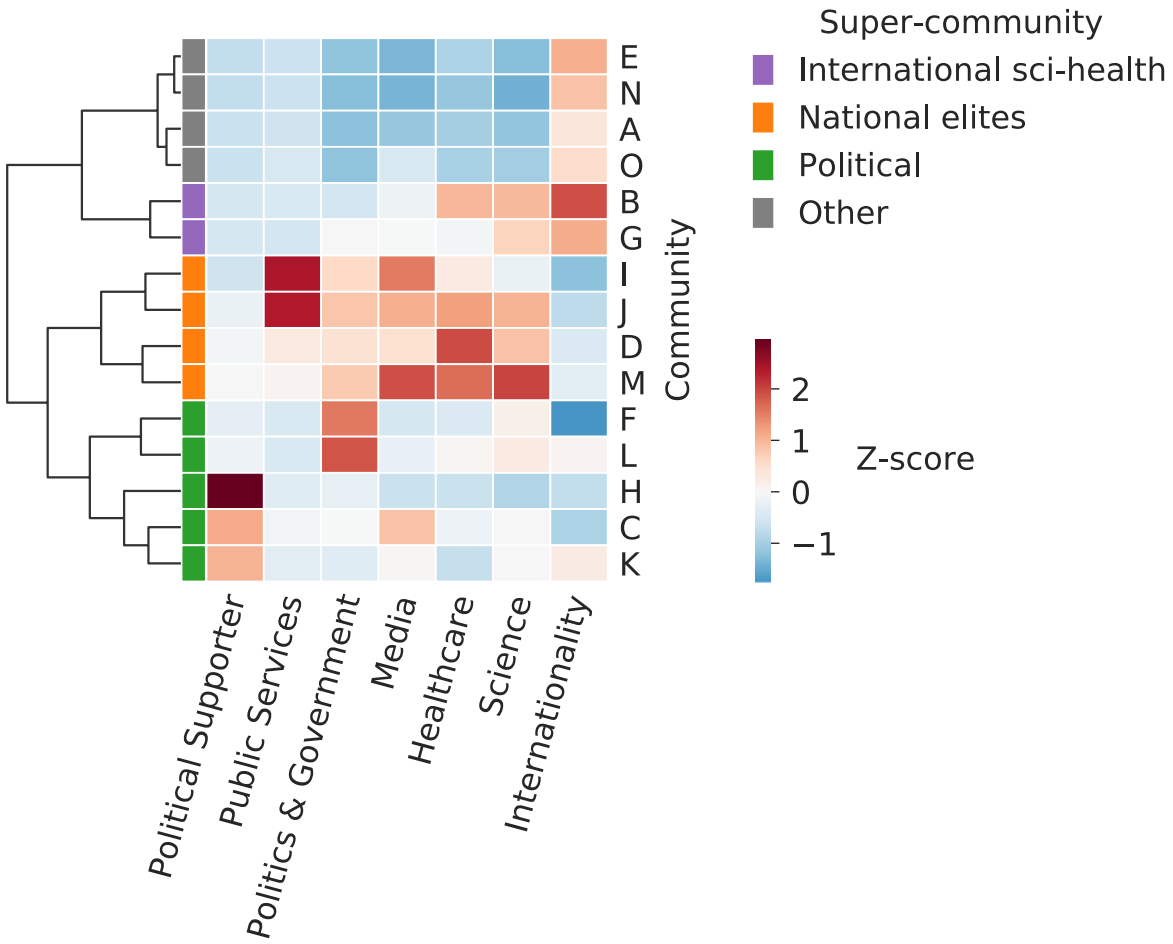


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Community structure

AI **analysis of tweet text** to characterize main topics within the communities

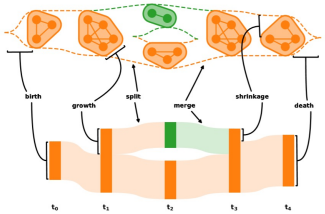
- strong association of topics and communities
- 4 super-communities

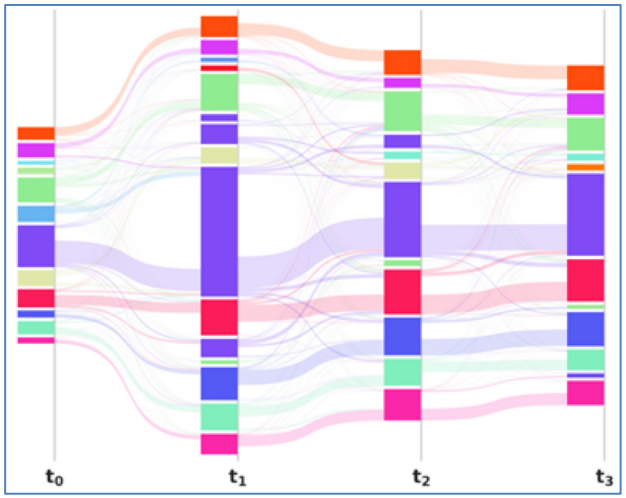


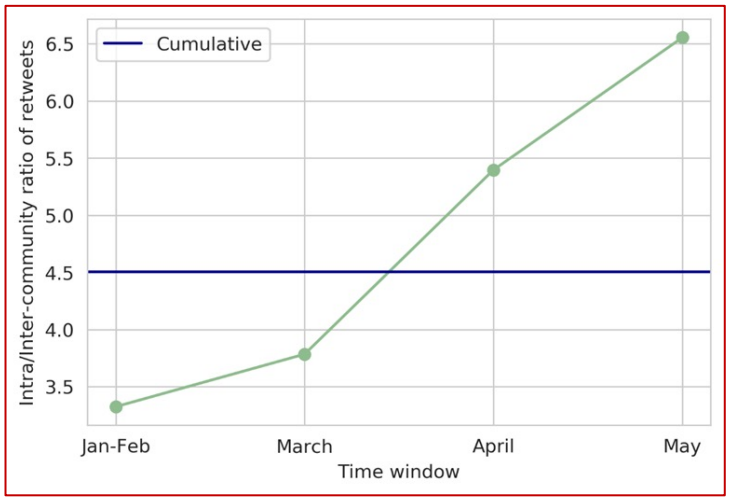
Time evolution: community ”ecology”

4 time windows (Jan/Feb, Mar, Apr, May: ”early” ”peak” ”late” phases of 1st COVID-19 wave)

Stable communities over time
Inter-cluster communication decreases over time
(community SEGREGATION, very typical of social networks)



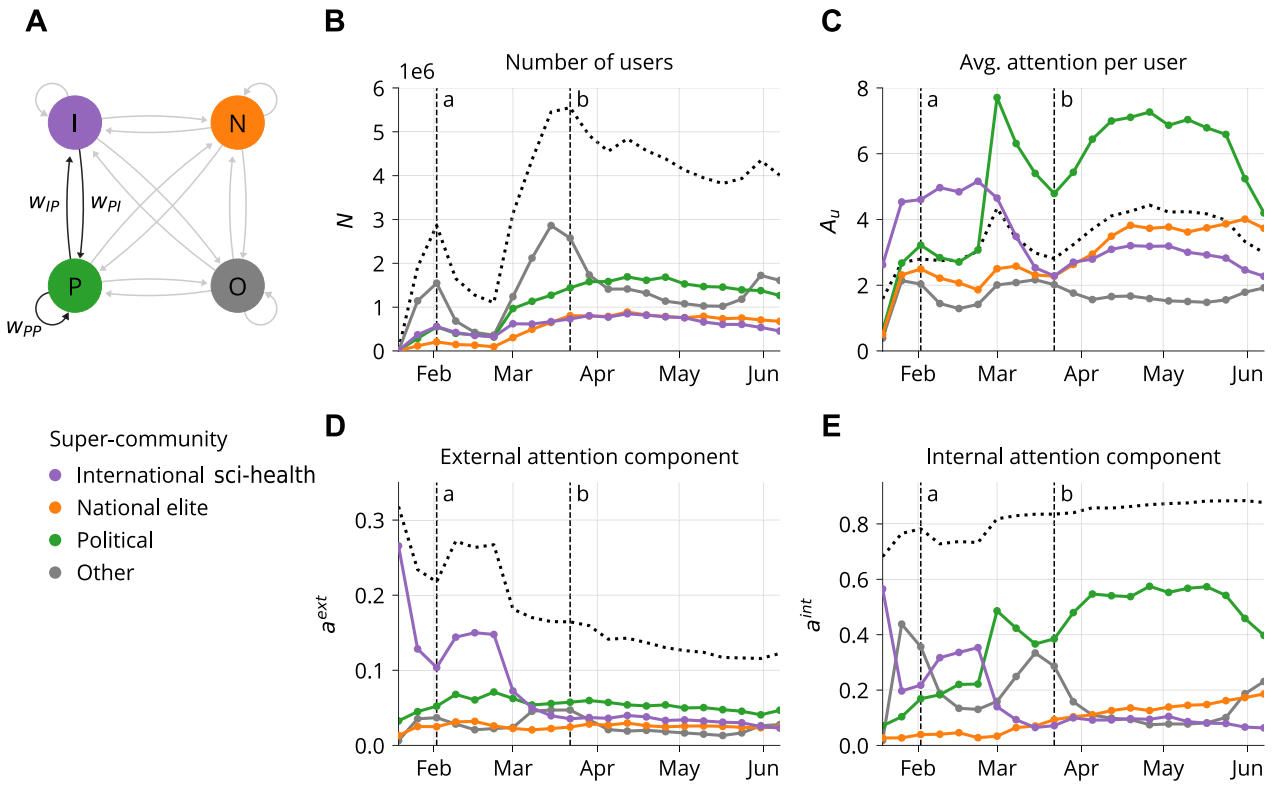




ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Network evolution: attention shift

Growing concern: increase of users talking about COVID-19
Initial phase: much "talk" in scientific community and attention to them
Scientific community loses attention over time, and political community activity increases
- Possibly the debate shifts from "technical" to "political"



AI approaches:
from "geno" to "pheno"

Protein sequence "embedding" in vector space

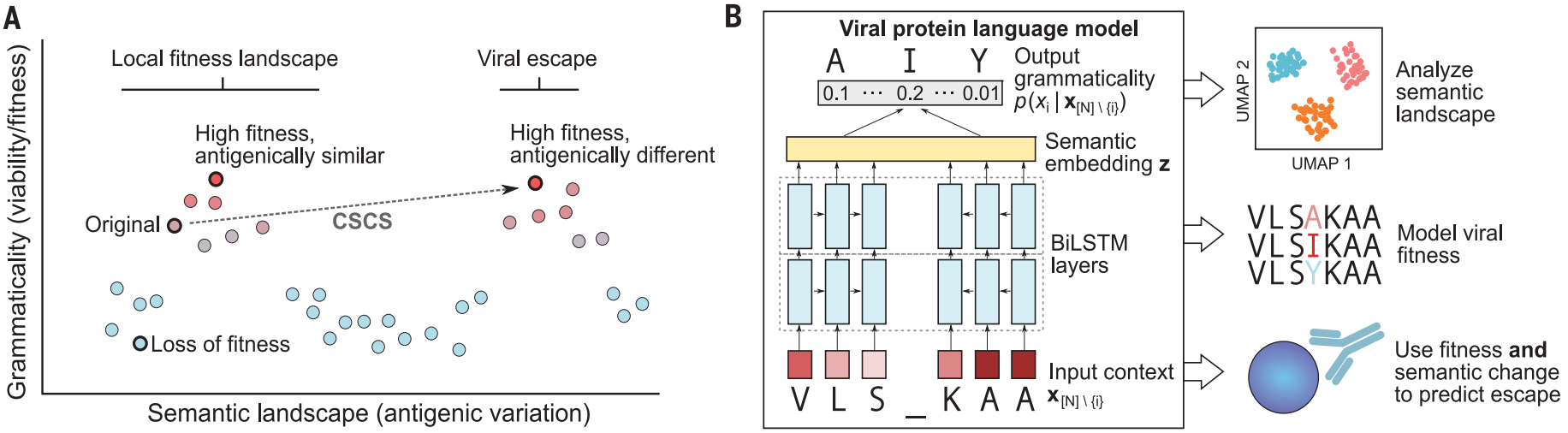
Application of text AI – NLP – to protein AA symbolic sequences:

Each protein sequence is "transformed" into an N-dim (1024) vector

Training: reconstruct sequence estimating missing aminoacid in the sequence

Calculate sequence "distances" (different from sequence alignment)

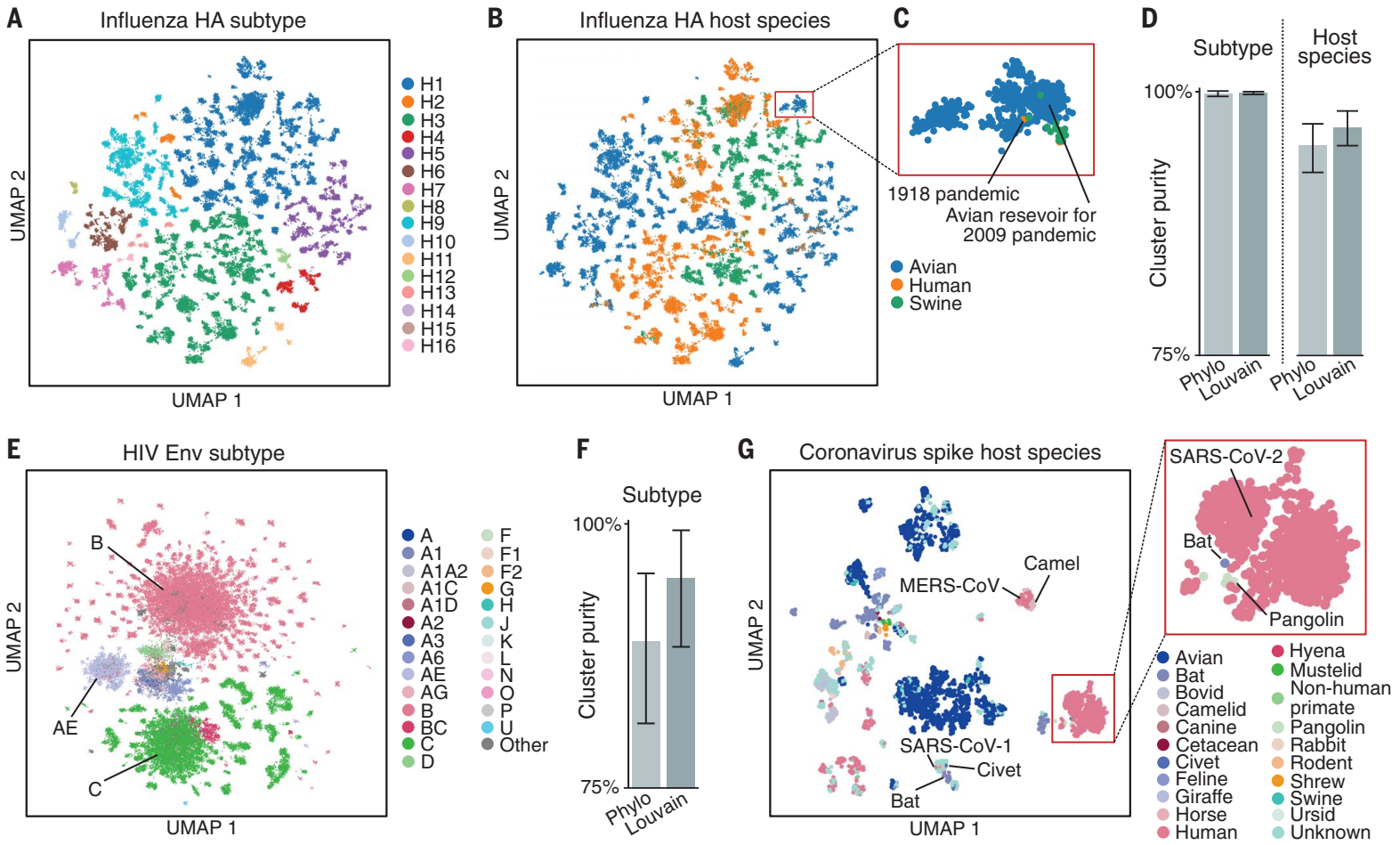
Grammaticality score: AA probability + "immune escape" (vector distance)



[Hie et al., Science 2021]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

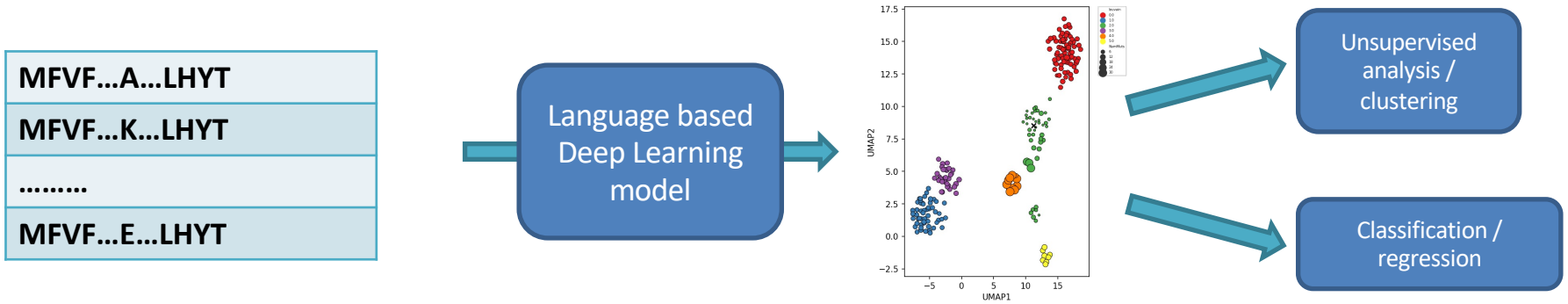
Dimensionality reduction: clustering by host



[Hie et al., Science 2021]

AI-based embedding

transform protein sequences into vectors (metrics)
these vectors can be used for ML/data analytics applications:
clustering & dimensionality reduction (visualization)
supervised classification & regression if **ground truth** available.



The classical paradigm is to compare protein variants through **phylogenetic distances** based on **sequence alignment and overlap**

Alignment vs embedding distance

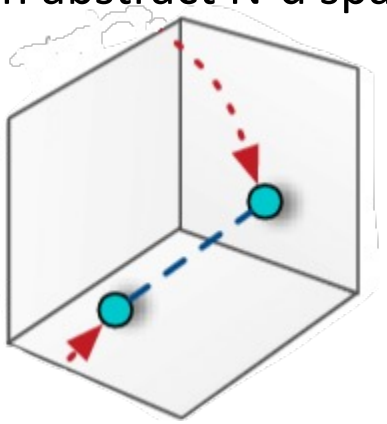
The **classical paradigm** is to compare protein variants through **phylogenetic distances** based on **sequence alignment and overlap**

Alignment-based (phylogenetic) distance: ratio of sequence overlap

```
A T A T G C T A G G C C A G C

T T A T G C T A T G C _ _ G C
```

Embedding distance: mathematical vector distance in abstract N-d space



Embedding distance **not (always) proportional** to sequence overlap. Even a small alignment distance (eg single AA substitution) can lead to a large embedding distance (depending on the "grammaticality" of the sequence)

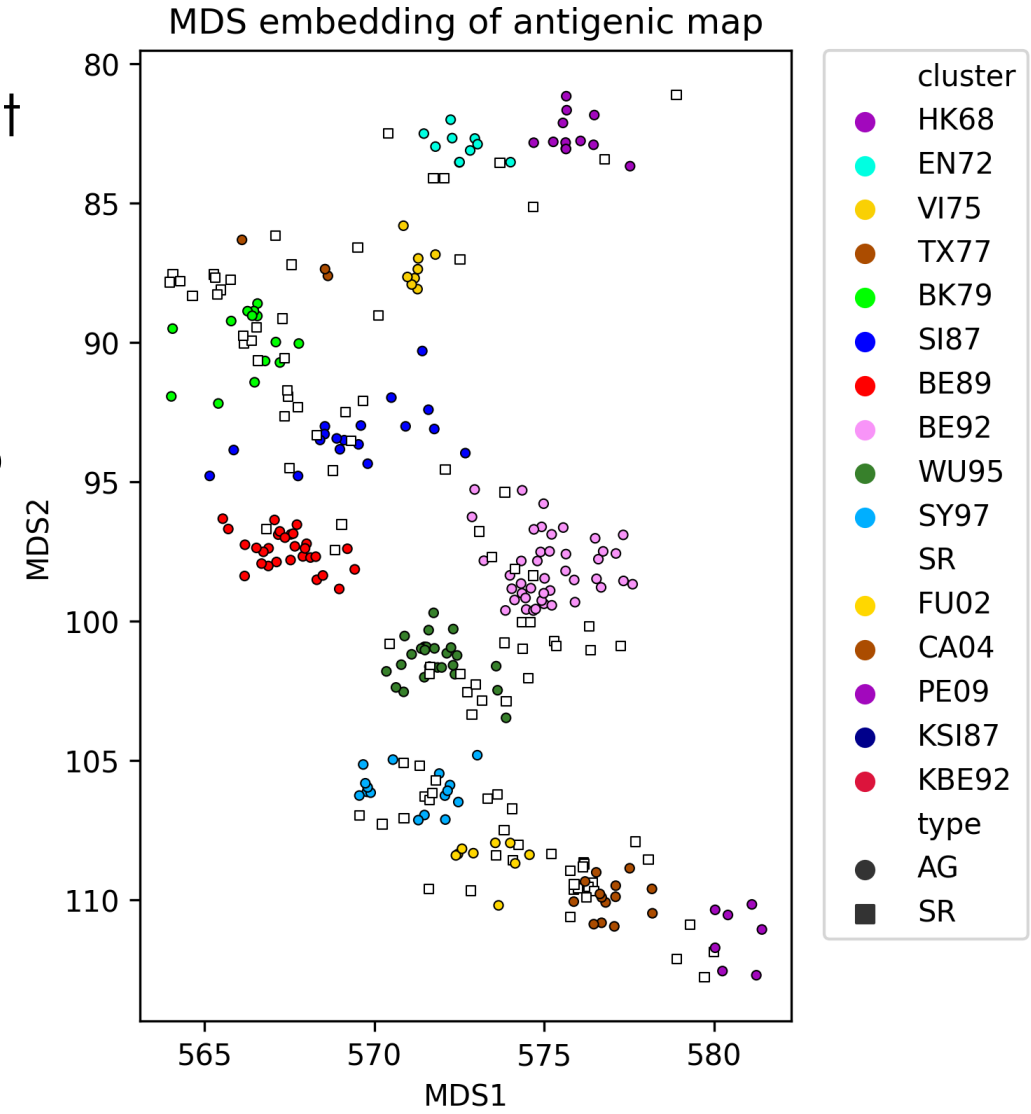
H3N2 HA1 HI assays: starting point

[Durazzi, Fouchier, Koopmans, Remondini Sci Rep 2025]

Antigenic map:
measure response of different sera to virus variants
[Smith et al Science 305 2004]

HI (Hemagglutinin Inhibition) assays represented as a **map** in 2-d space

OUR AIM: reproduce antigenic maps from protein sequence alone through protein AI embedding



H3N2 HA1 domain embedding with NLP

NLP Natural Language Processing (AI text tools)

"Words" = Aminoacids

"Sentence" = Protein

Learning the **language** of HA1:

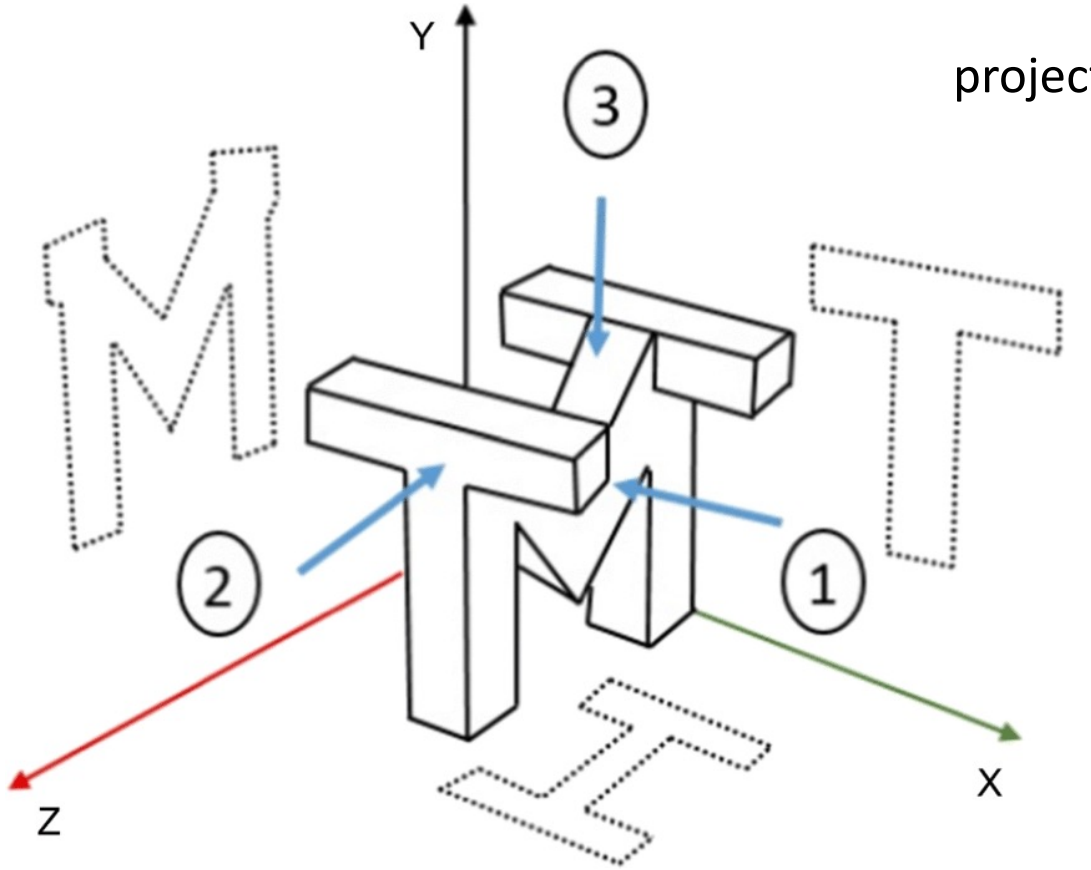
biLSTM Recursive Network trained on approx 40k HA1 sequences *[Hie et al., Science 371 2021]*

NOTE: same results with protBERT Deep Learning model

Comparison with:

- phylogenetic protein comparison
- ML based on physicochemical AA properties

Extracting info from embedding: regression to ground truth



projection: 1024d to 2d

Each HA1 sequence becomes a point in 1024-d space

Is antigenic map info contained in this embedding?

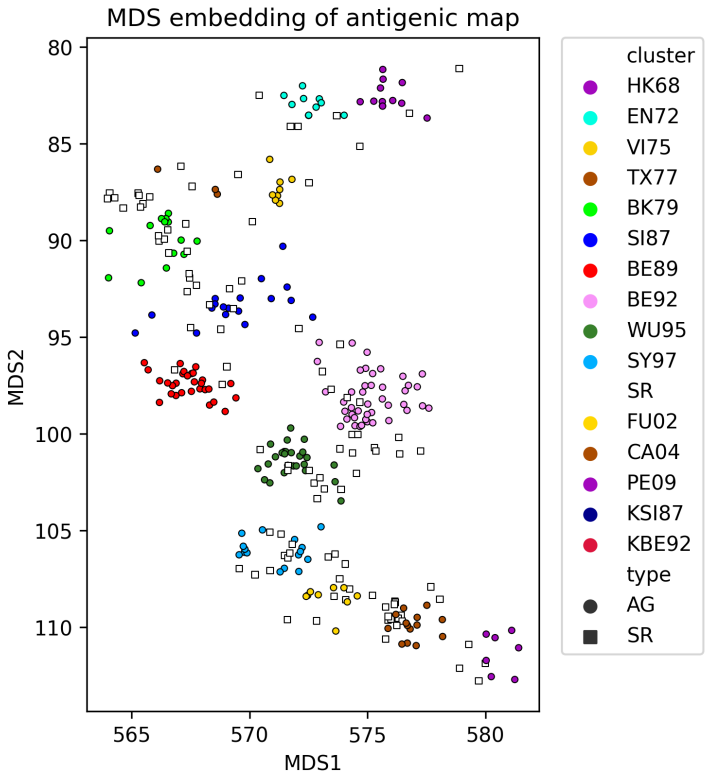
Geno-to-pheno: from HA sequences to AM

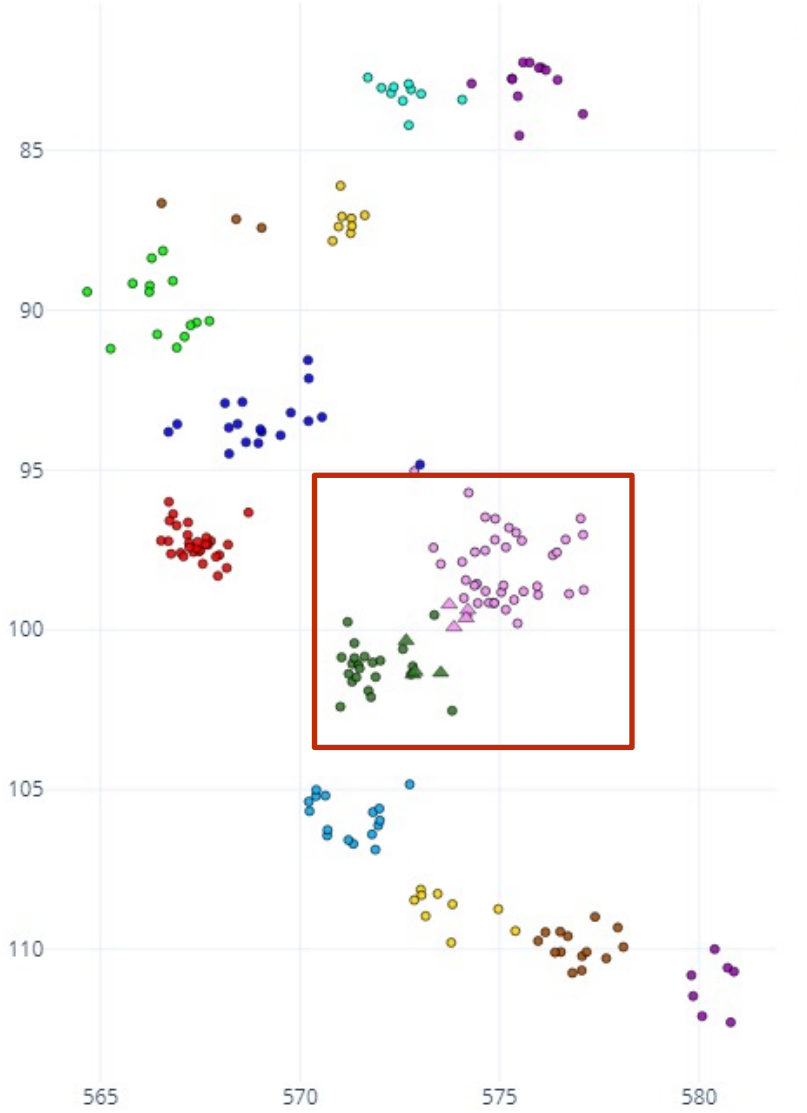
Linear (ridge) regression

Train error: 0.54 a.u.

Test error: 0.92 a.u.

Same cluster structure, also **single-AA** BE92-WU95 change



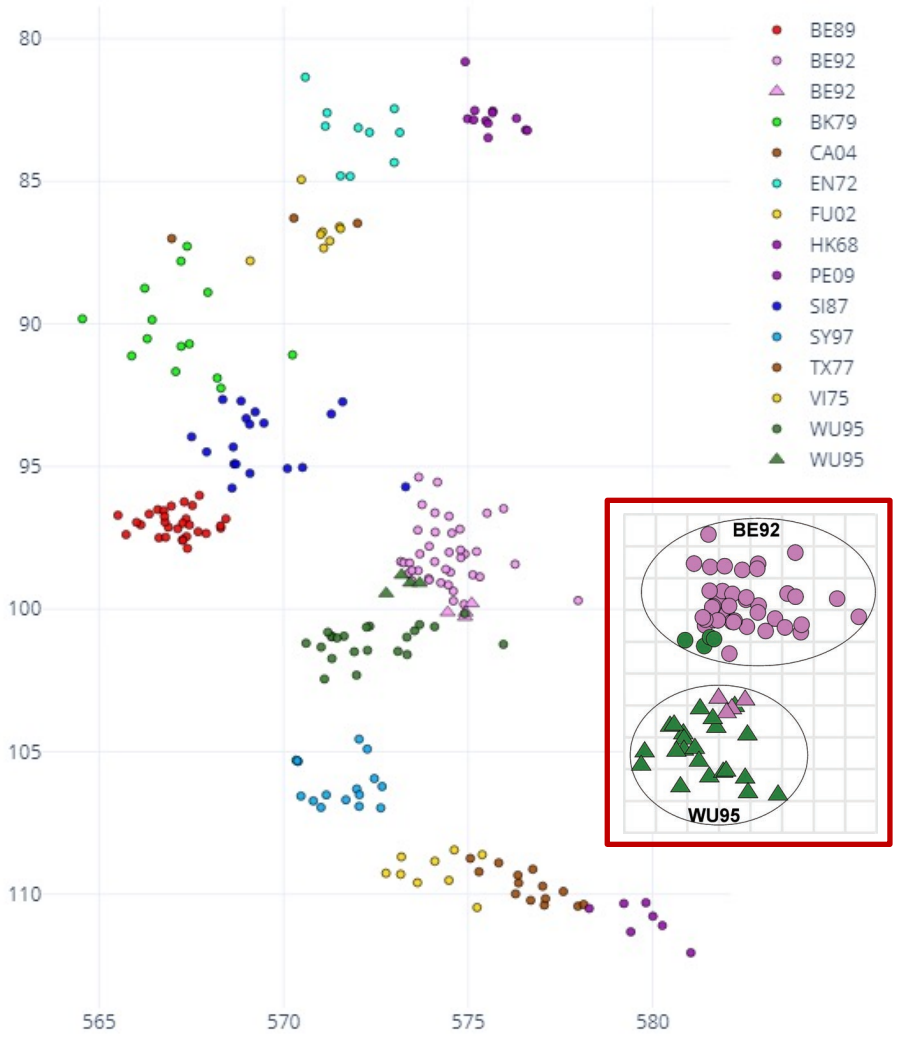


Phylogenetic distance regression

1024D MDS of Hamming distance between protein sequences
+
Ridge regression

Train error: 0.83 a.u.
Test error: 1.10 a.u.

Single AA cluster shift **not** recovered



CHV signature

(3x329=) 987D vectors of Charge-Volume-Hidropathy values for each AA

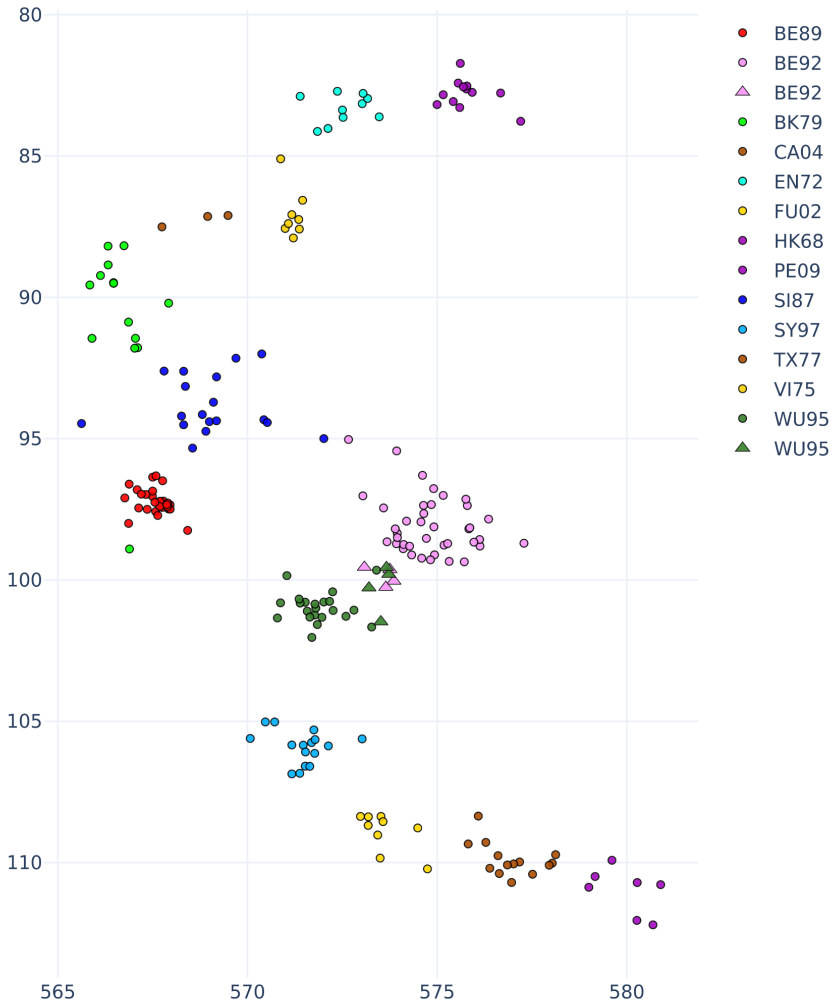
+

Ridge regression

Train error: 0.45 a.u.

Test error: 1.31 a.u.

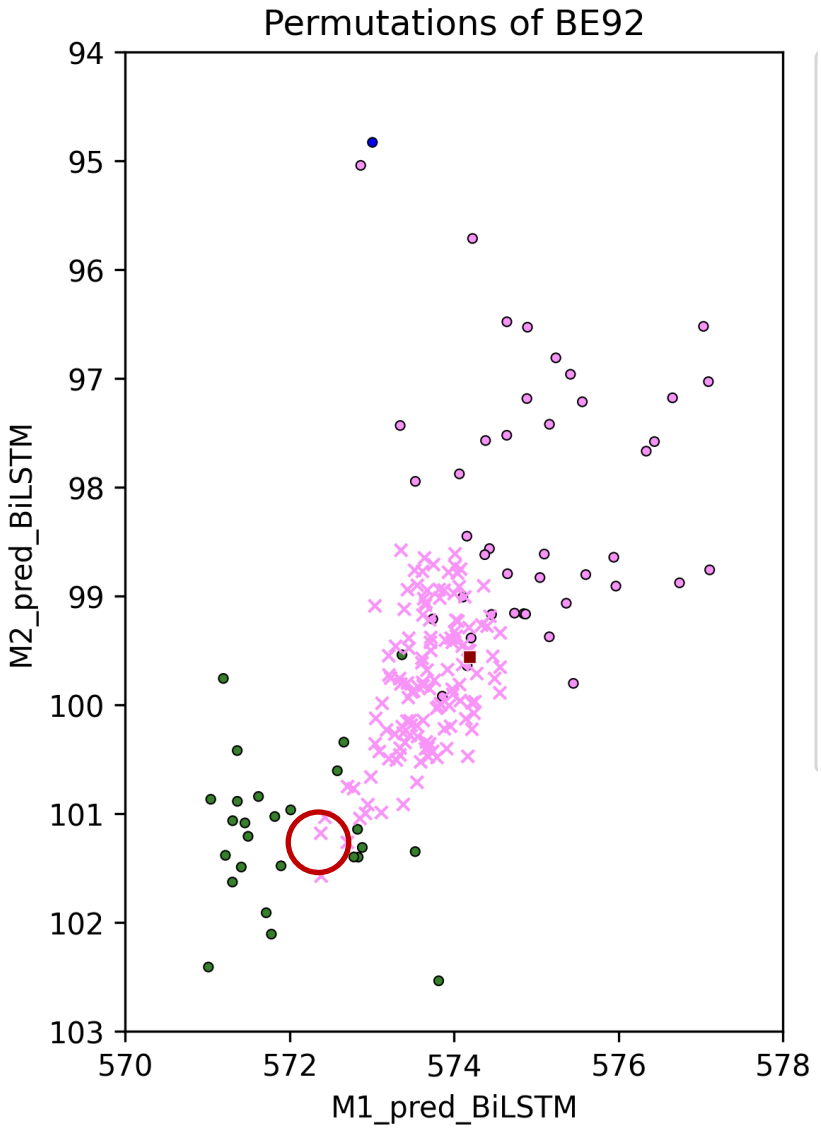
Single-AA shift **not** recovered



In-silico Deep Mutational Scan experiment

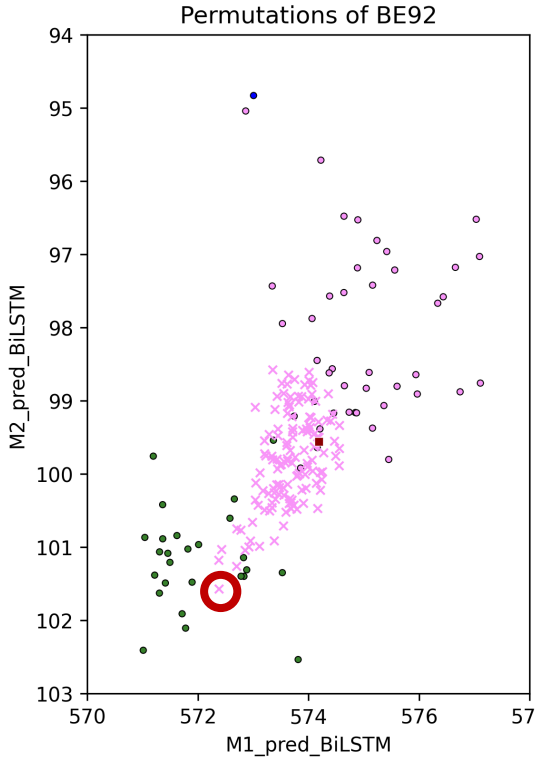
start from a **real sequence** in a cluster and **predict** the **AM coordinates** of the 20 **AA substitutions** at specific sites

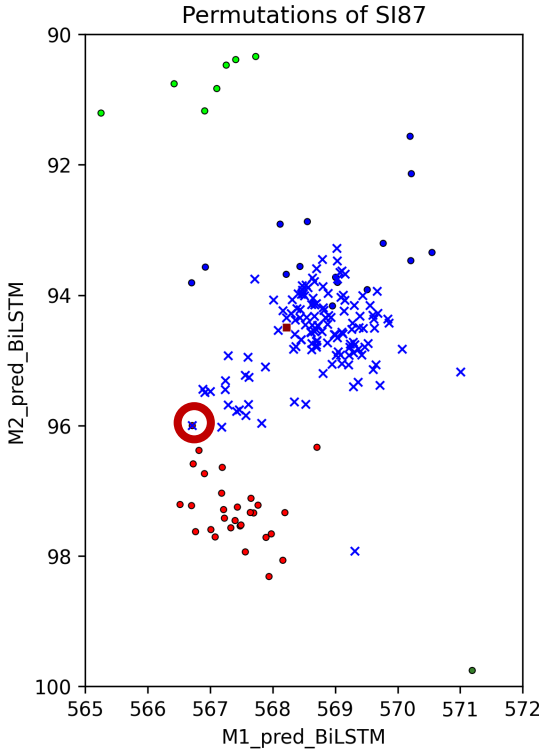
- rank the 140 substitutions (7 sites) combining **grammaticality** and **antigenic distance** from the sequence of origin



BiLSTM embedding

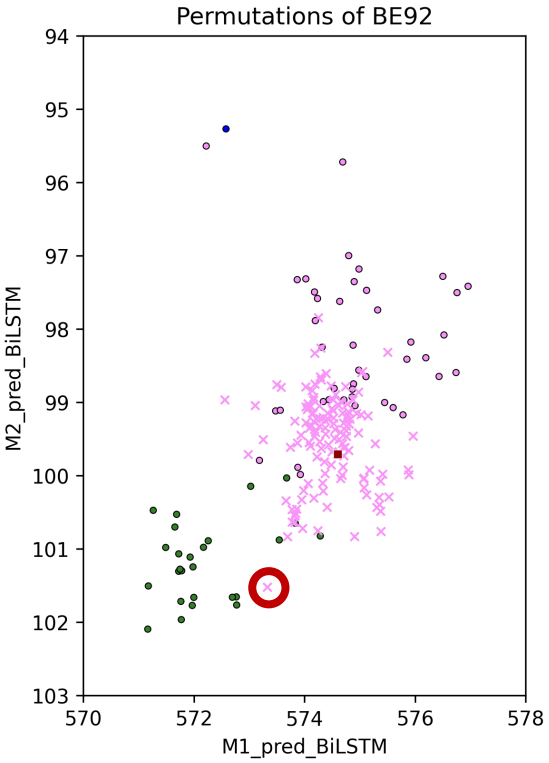
HK68: T155Y rank=86
EN72: Q189K rank=6
VI75: G158E rank=2 and D193N rank=7
TX77: K156E rank=56
BK79: Y155H rank=12
BK79: S159Y rank=9
BK79: K189R rank=15
SI87: N145K rank=1 and E156K rank=16
BE92: N145K rank=1
WU95: K156Q rank=14 and E158K rank=54
SY97: Q156H rank=17
FU02: K145N rank=9
CA04: K158N rank=2 and N189K rank=1
Avg. Rank=18

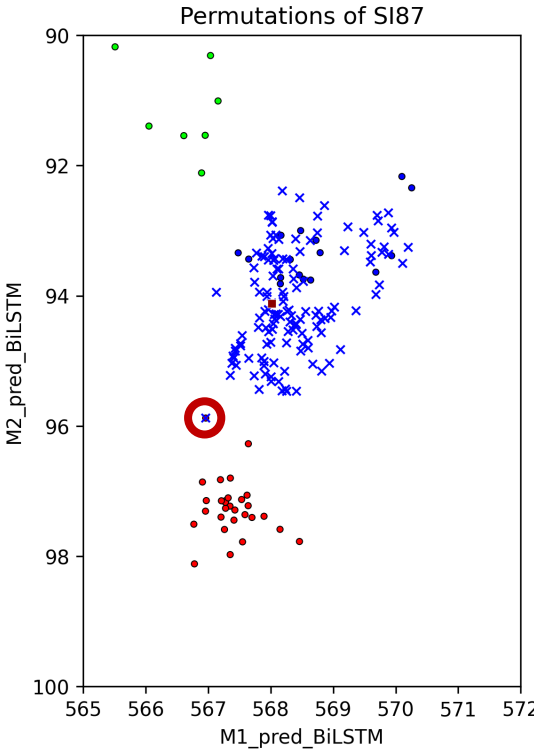




ProtBERT embedding

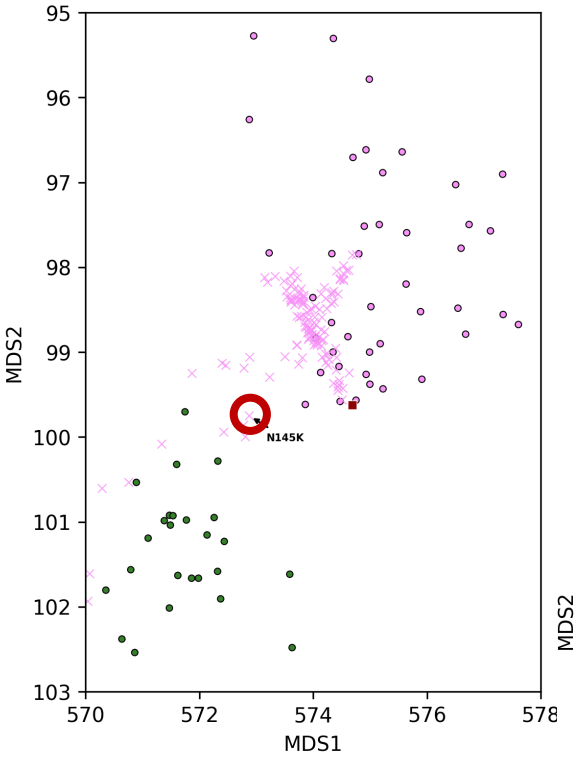
HK68: T155Y rank=8
EN72: Q189K rank=2
VI75: G158E rank=4 and D193N rank=5
TX77: K156E rank=26
BK79: Y155H rank=7
BK79: S159Y rank=9
BK79: K189R rank=13
SI87: N145K rank=1 and E156K rank=12
BE92: N145K rank=1
WU95: K156Q rank=34 and E158K rank=29
SY97: Q156H rank=35
FU02: K145N rank=4
CA04: K158N rank=12 and N189K rank=1
Avg. Rank=12

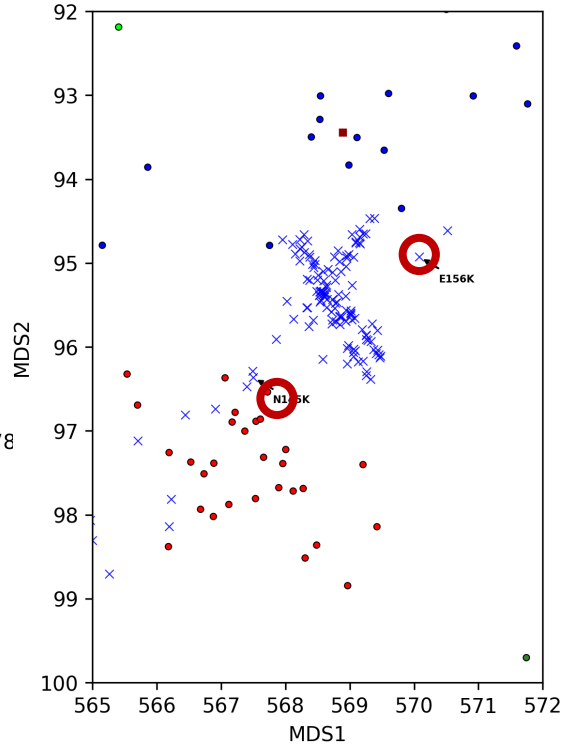




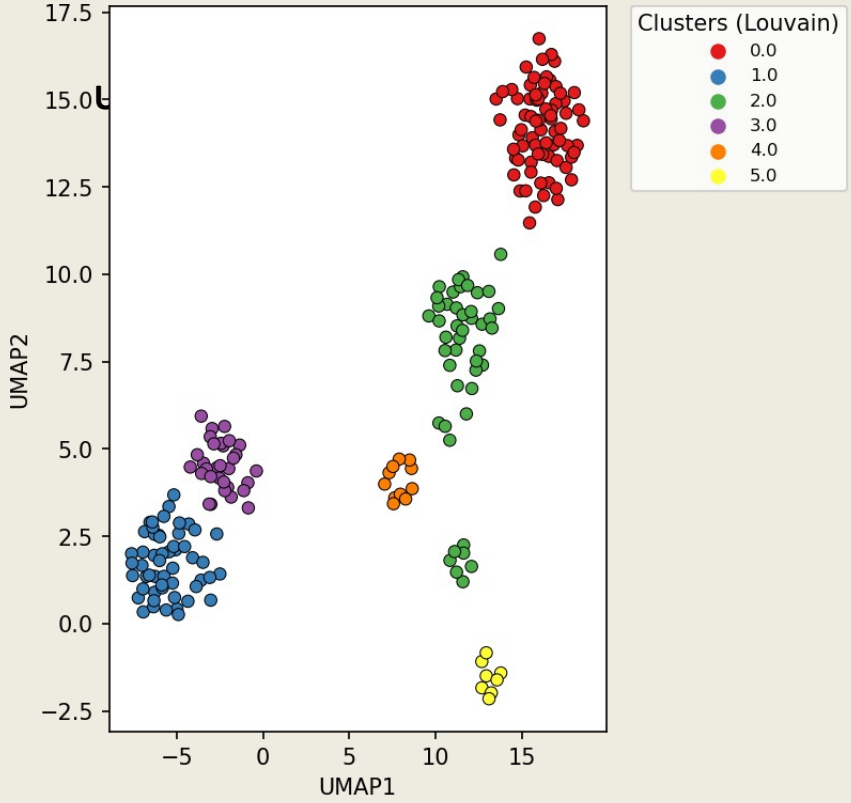
CVH embedding

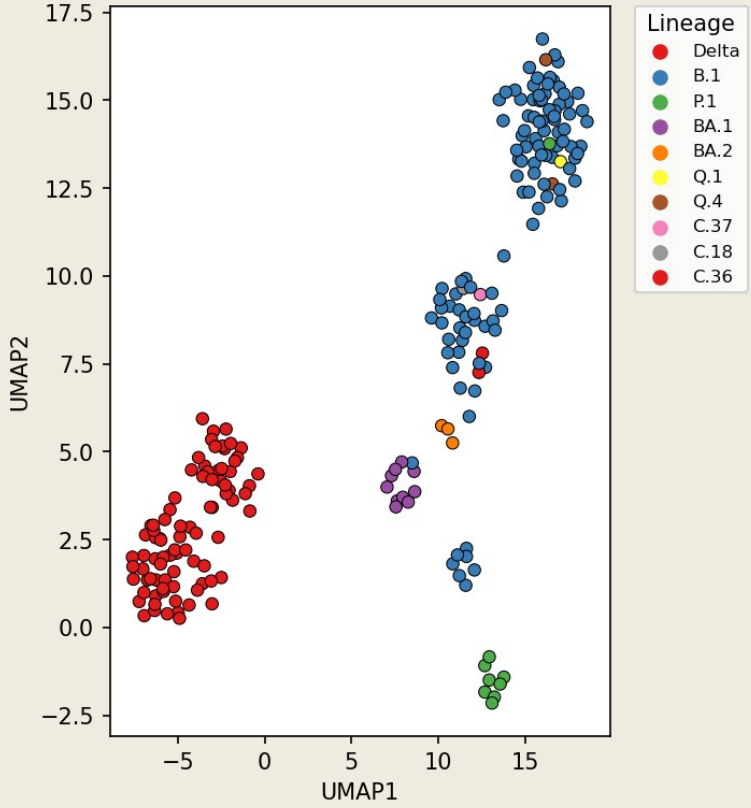
HK68: T155Y rank=104
EN72: Q189K rank=97
VI75: G158E rank=28 and D193N rank=123
TX77: K156E rank=106
BK79: Y155H rank=58
BK79: S159Y rank=68
BK79: K189R rank=106
SI87: N145K rank=30 and E156K rank=73
BE92: N145K rank=23
WU95: K156Q rank=139 and E158K rank=134
SY97: Q156H rank=122
FU02: K145N rank=29
CA04: K158N rank=71 and N189K rank=108
Avg. Rank=83 (median=97)





Other examples: SARS-COV2 GENOTYPING IN BOLOGNA





Good overlap between the clusters and lineages

One possible interesting cluster jump corrected a labeling error

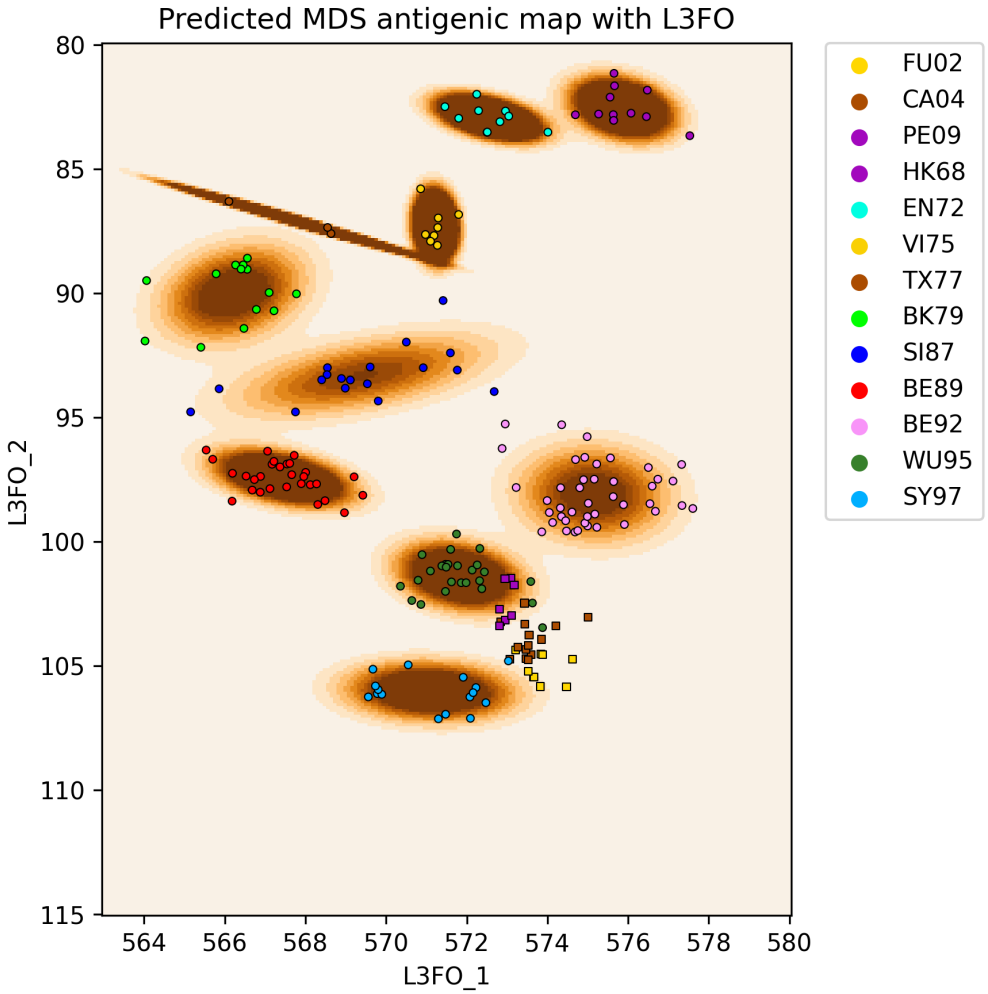
Leave-the future-cluster-out: BiLSTM

Remove last N clusters in time from the map, and try to predict if they would result as outliers with respect to the previous ones

BiLSTM embedding

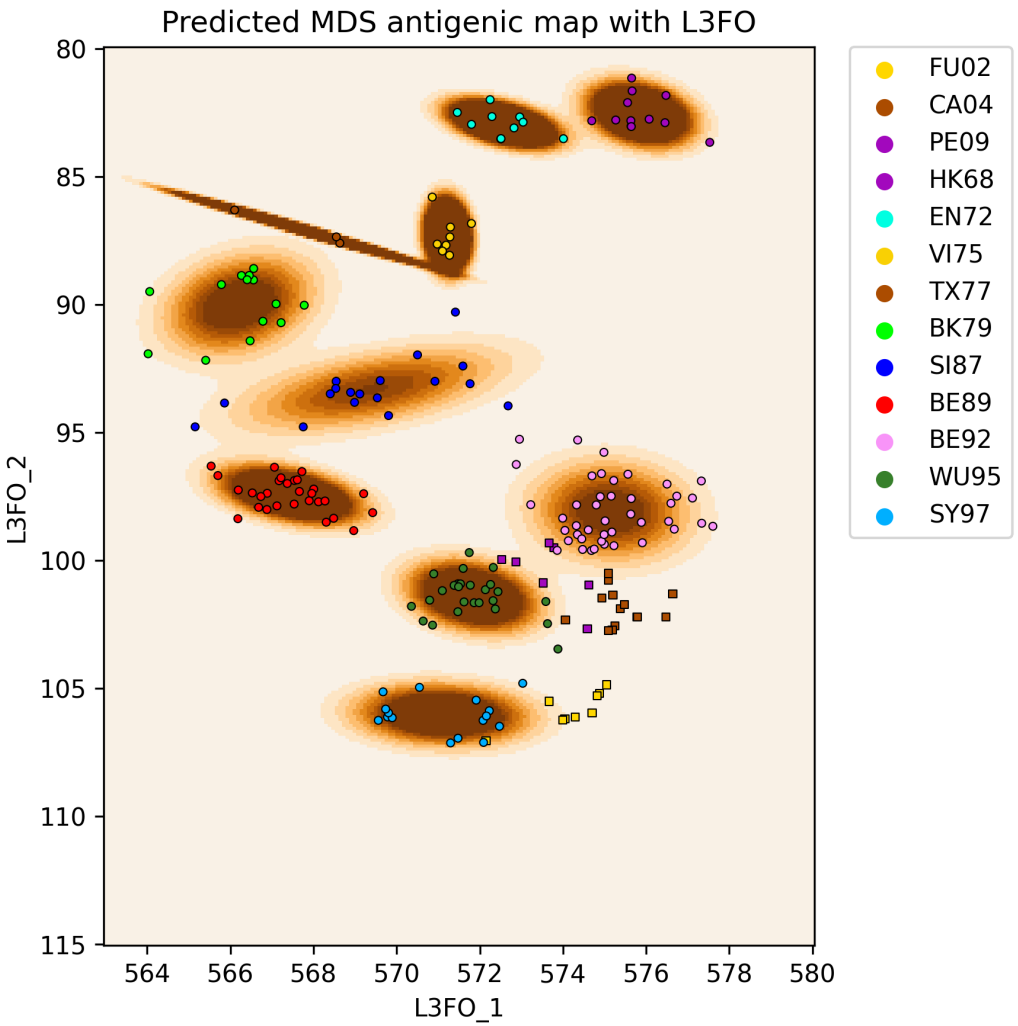
how many samples are correctly predicted outside the last training cluster in time.

L1FO: 7/7 predicted outside (100%)
L2FO: 12/21 (57%)
L3FO: 26/30 (87%)
L4FO: 44/46 (95%)
L5FO: 8/72 (11%)
L6FO: 115/115 (100%) but 114 are in the second-last cluster



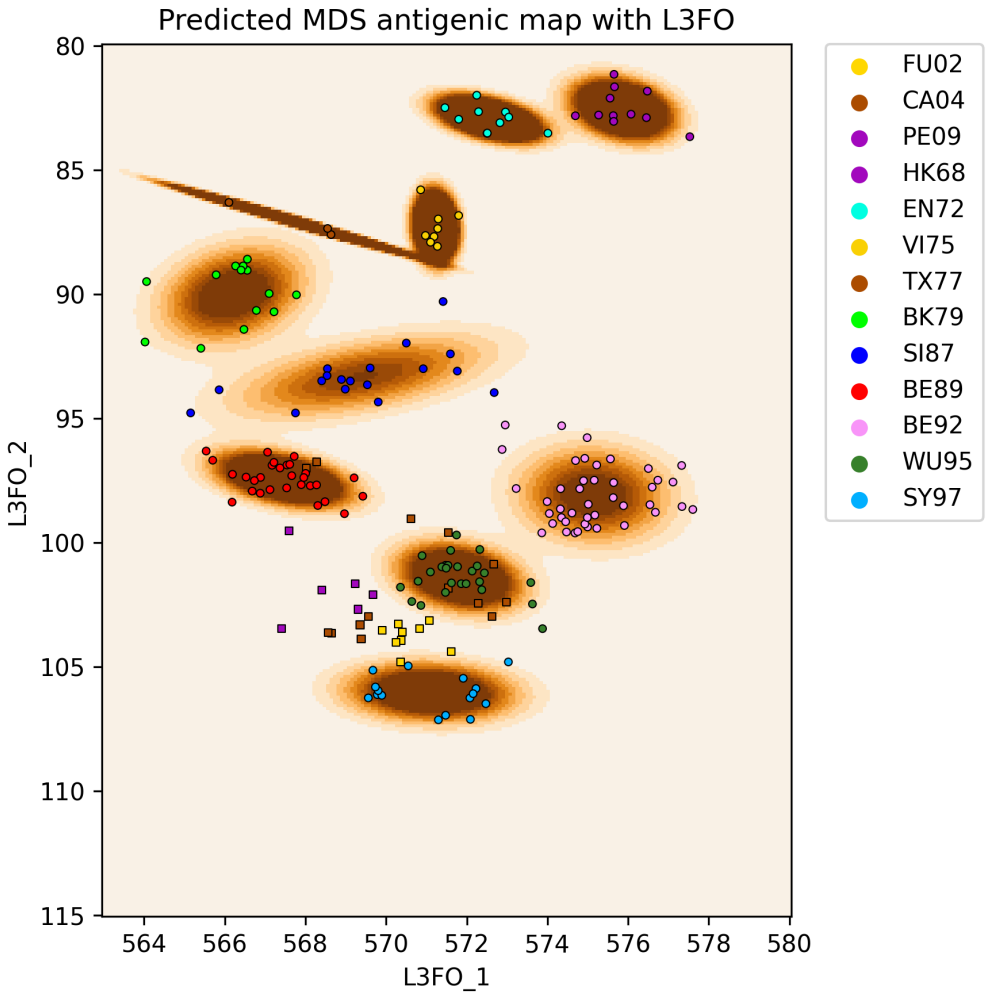
ProtBERT embedding

L1FO: 6/7 predicted outside (86%)
L2FO: 21/21 (100%)
L3FO: 26/30 (87%)
L4FO: 33/46 (72%)
L5FO: 31/72 (43%)
L6FO: 114/115 (99%) but 50 are in the second-last cluster



CVH embedding

L1FO: 4/7 predicted outside (57%)
L2FO: 20/21 (95%)
L3FO: 28/30 (93%)
L4FO: 42/46 (91%)
L5FO: 44/72 (61%)
L6FO: 114/115 (100%) but 82 are in the second-last cluster

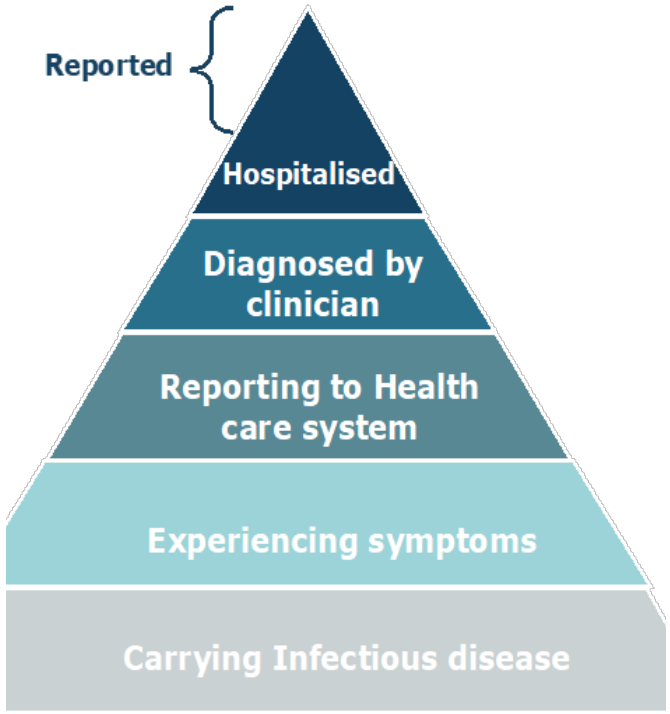


Combining experiments and models in a "One health approach"

One Health initiative: VEO case study

Bologna metropolitan area monitoring:

- Microbiological data: urban wastewater microorganisms
- Clinical data: COVID, salmonella, campylobacter, shigella, ... cases
- Socioeconomic data: human mobility and traffic
- Demographic data: comorbidities, vaccinations
- Weather and climate data: rainfall, humidity, temperatures, pollution
- Veterinary data: pets, livestock production, wild animals
- Social data: monitoring Twitter on the keywords 'COVID' and 'vaccines'



3-year monitoring of COVID-19 in Bologna metropolitan area 2021-2023

Epidemiological mathematical model adjusted on clinical data

Sars-CoV-2 in urban sewage

SARS-CoV-2 genomic tracing of lineage evolution over time

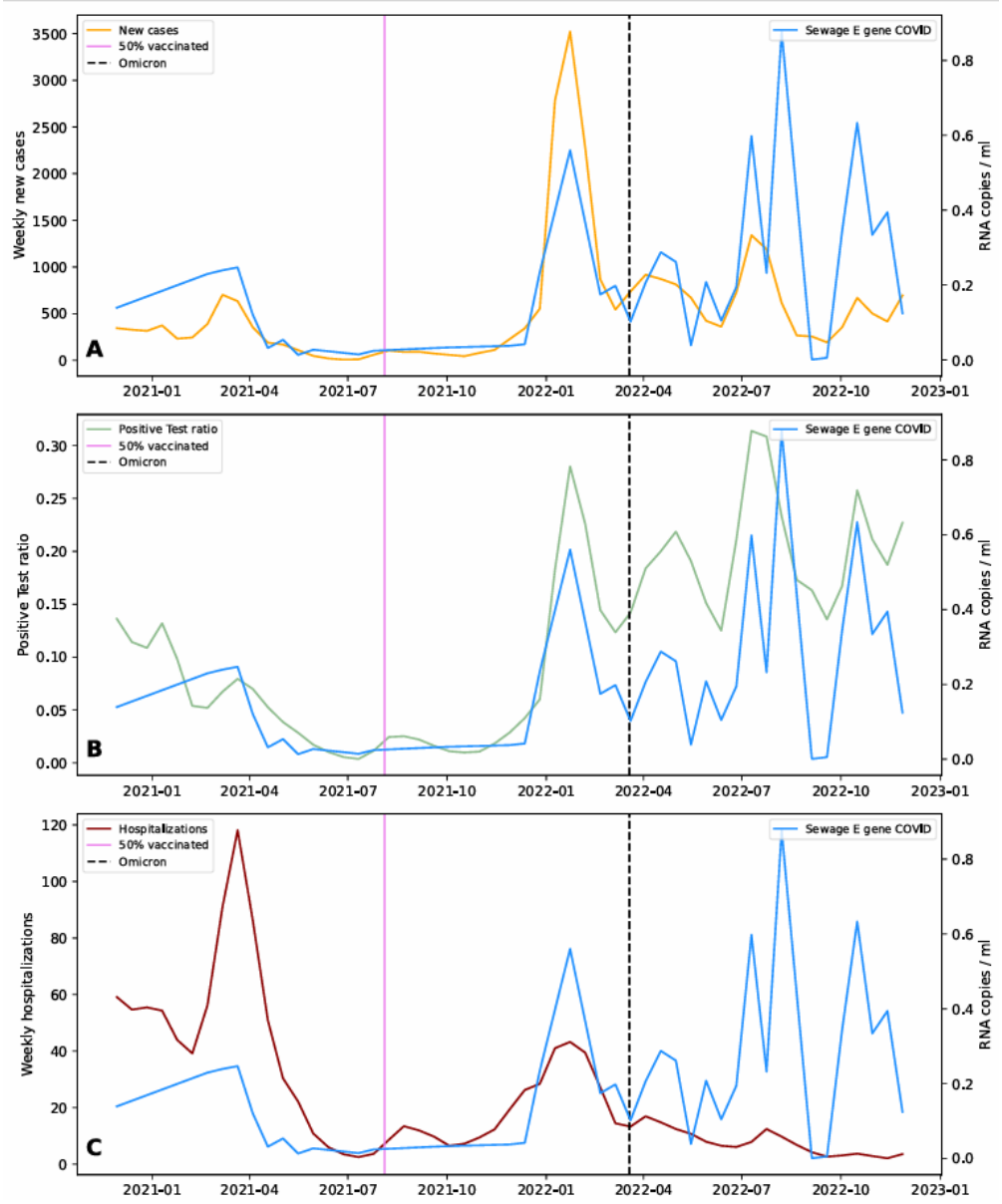
Road traffic time series

Vaccination coverage of the population

HERA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Sars-Cov-2 in wastewater + cases



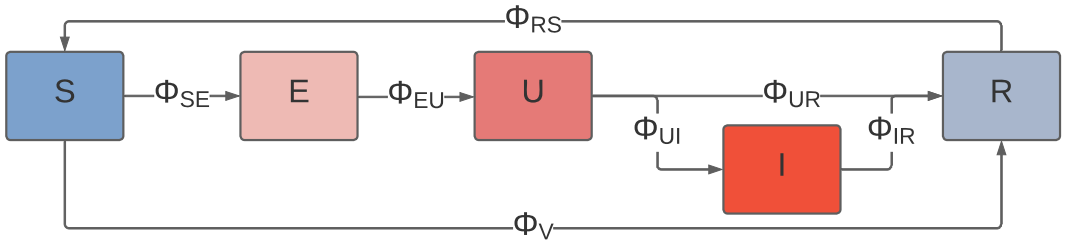
last phase: hospitalizations & new cases decline even when viral load increases (r = 0.73)

why?

The model: epidemics evolution

(with Prof A. Bazzani, G. Colombini, E. Lunedei et al.)
ODE system to describe epidemics spread

$$
\begin{aligned}
\dot{S}(t) &= -\Phi_{S\to E}(t) - v(t)S(t) + v(t-T_R)S(t-T_R) \\
&\quad + (1-\alpha)\int_0^\infty \Phi_{S\to E}(t-T_E-T_R-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad + \alpha\int_0^\infty \Phi_{S\to E}(t-T_E-T_I-T_R-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
\dot{E}(t) &= \Phi_{S\to E}(t) - \Phi_{S\to E}(t-T_E) \\
\dot{U}(t) &= \Phi_{S\to E}(t-T_E) - \int_0^\infty \Phi_{S\to E}(t-T_E-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
\dot{I}(t) &= \alpha\int_0^\infty \Phi_{S\to E}(t-T_E-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad - \alpha\int_0^\infty \Phi_{S\to E}(t-T_E-T_I-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
\dot{R}(t) &= (1-\alpha)\int_0^\infty \Phi_{S\to E}(t-T_E-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad - (1-\alpha)\int_0^\infty \Phi_{S\to E}(t-T_E-T_R-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad + \alpha\int_0^\infty \Phi_{S\to E}(t-T_E-T_I-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad - \alpha\int_0^\infty \Phi_{S\to E}(t-T_E-T_I-T_R-\tau)\rho(\tau;T_U,\sigma_U)\,d\tau \\
&\quad + v(t)S(t) - v(t-T_R)S(t-T_R)
\end{aligned}
$$



Susceptible
Exposed
Unreported + Isolated
Recovered

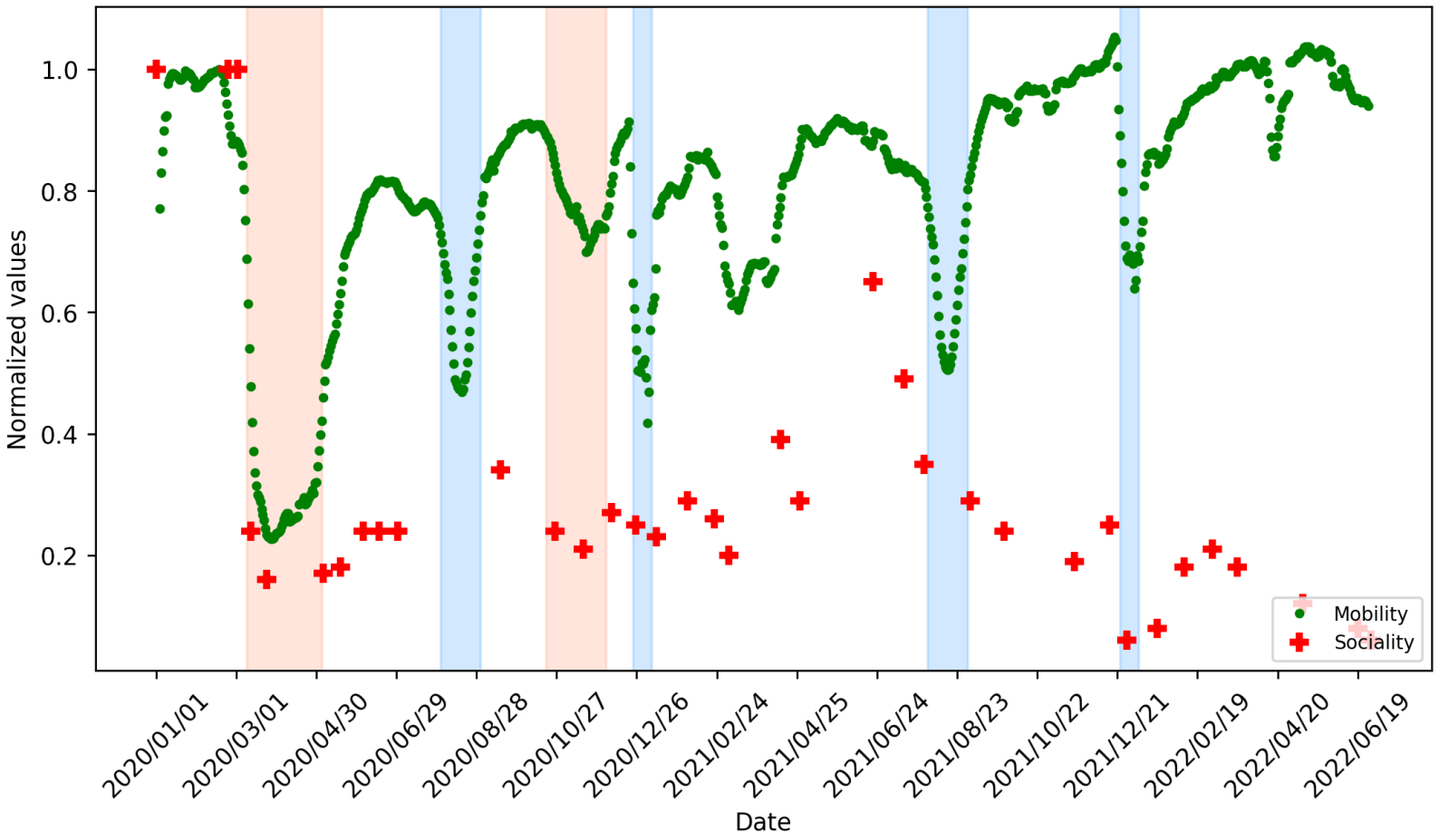
Virus spread = virus (intrinsic) **infectiveness**
+ human **sociability** (contacts)

Sociability and mobility

Sociability parameter: estimated on a weekly basis to fit the model output (#hospitalizations) with the clinical observations

Mobility: measures of road traffic in Bologna metropolitan area

Sociability and mobility: relations



Pink regions: lockdowns and curfews

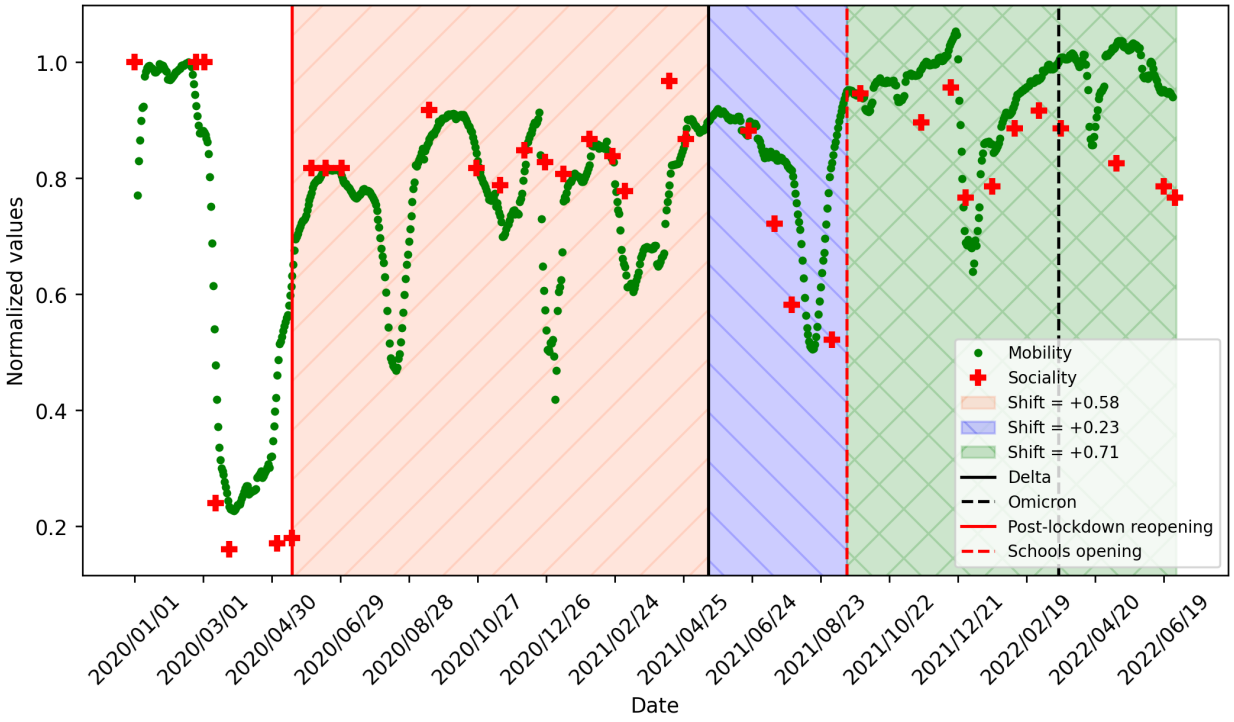
Blue regions: holidays

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Mobility as a proxy for sociability

Additive shift to re-normalize mobility and sociability

- High correlation (r=0.76)
- **Mobility can be used as a proxy to parametrize sociability** in the model for *short* periods (2-3 M)



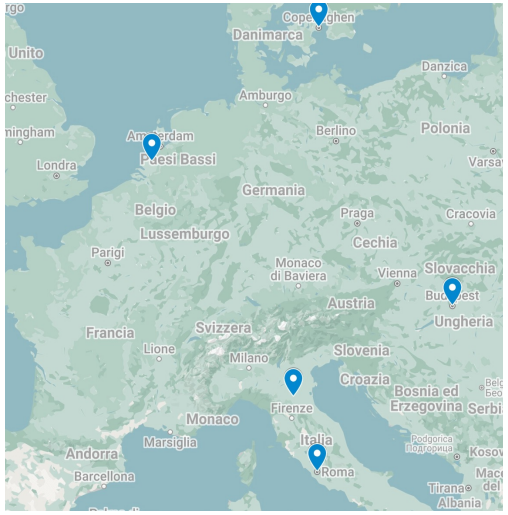
- **Shifts related to difference between protected and unprotected social interactions**
- Negligible shift at first phase (still not much protection) and summer 2021
- Larger shifts during periods of increased sensitivity to control measures (distancing, facial masks)

Day	Event	Shift
18/05/2020	Activities reopening (bar, restaurants)	0.58
17/05/2021	Delta variant in Emilia Romagna	0.23
15/09/2021	Schools reopening	0.71

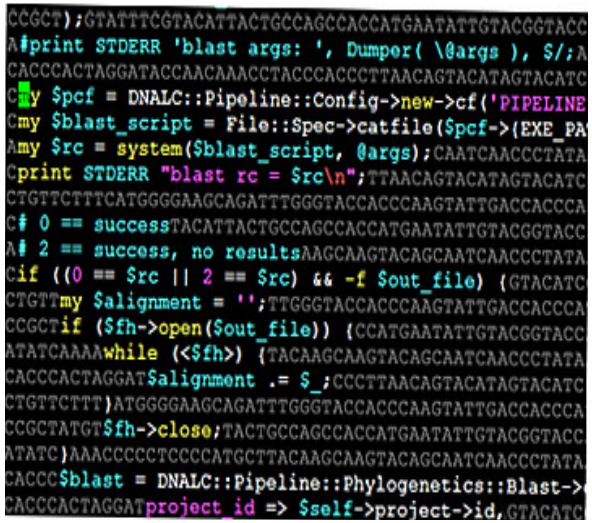
Mathematical & network-based approaches to metagenomic data of wastewater samples

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

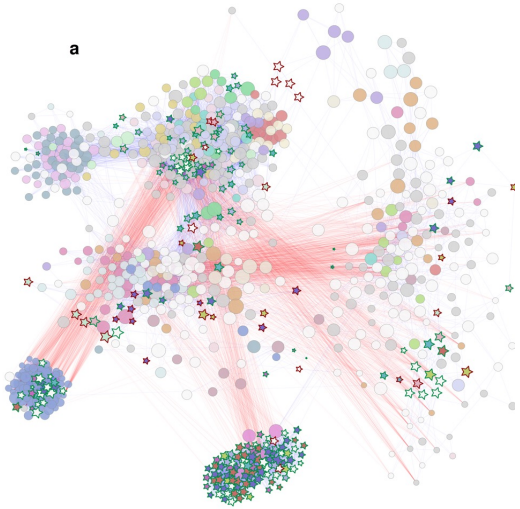
Activities within VEO



European Wastewater Cities Project



Bioinformatic Pipelines for high-volume, complex metagenomics

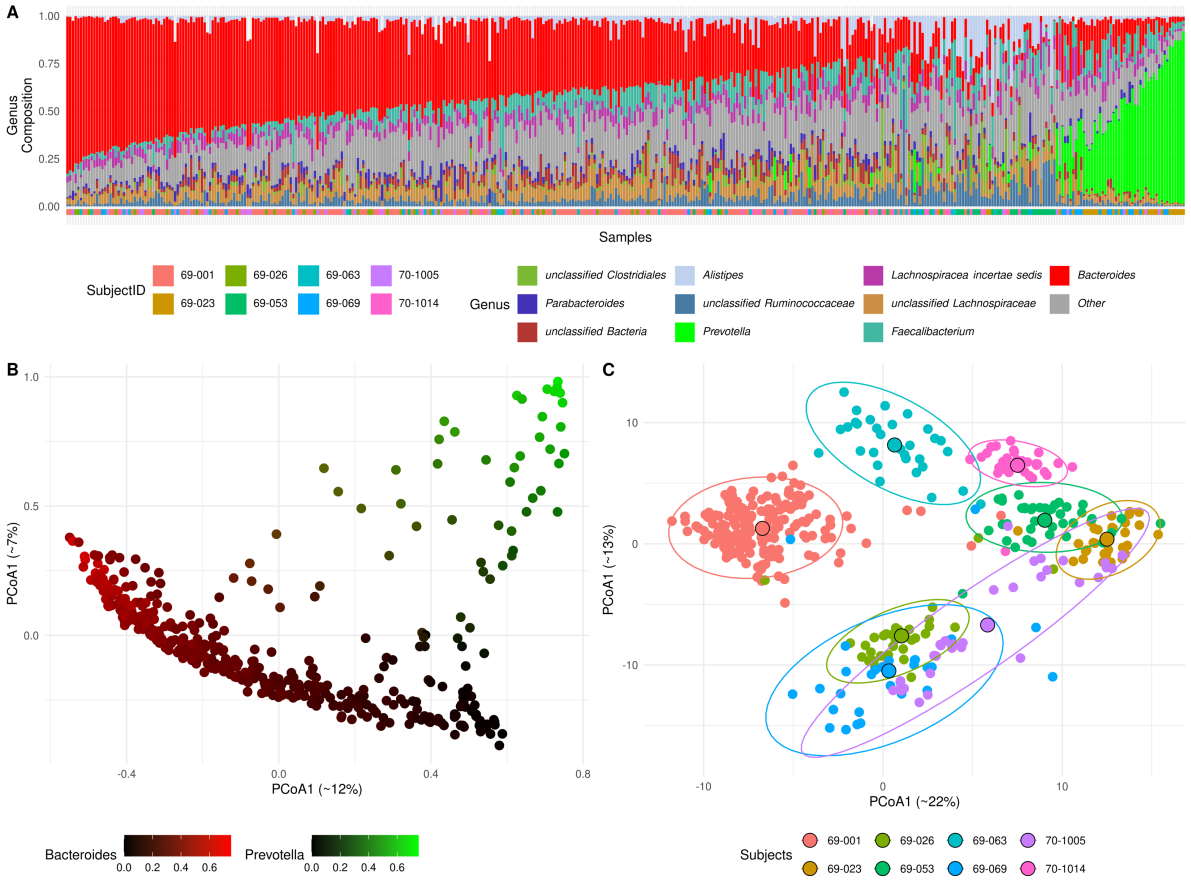


Tailored Data Analysis Statistical and network approaches for NGS

Times series (2y) of metagenomic (bacterial population) data in 7 Eu cities

From Physics to Compositionality: A Metric Dilemma

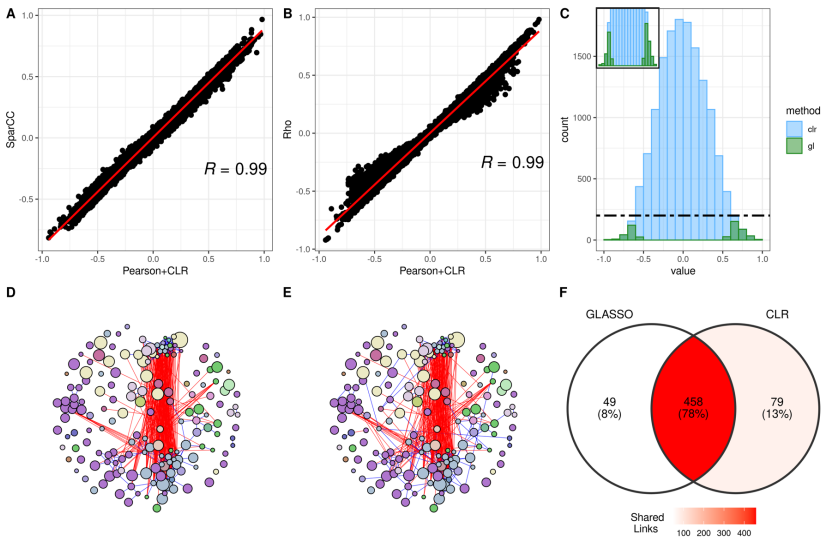
Metagenomics data are **compositional**: positive and constant-sum (like pdf)

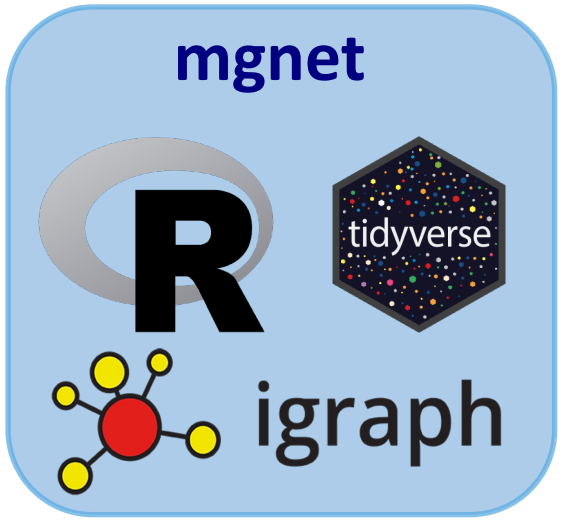


Bray–Curtis metrics (left) emphasizes dominant species

Aitchison metrics (right) reveals subject stratification

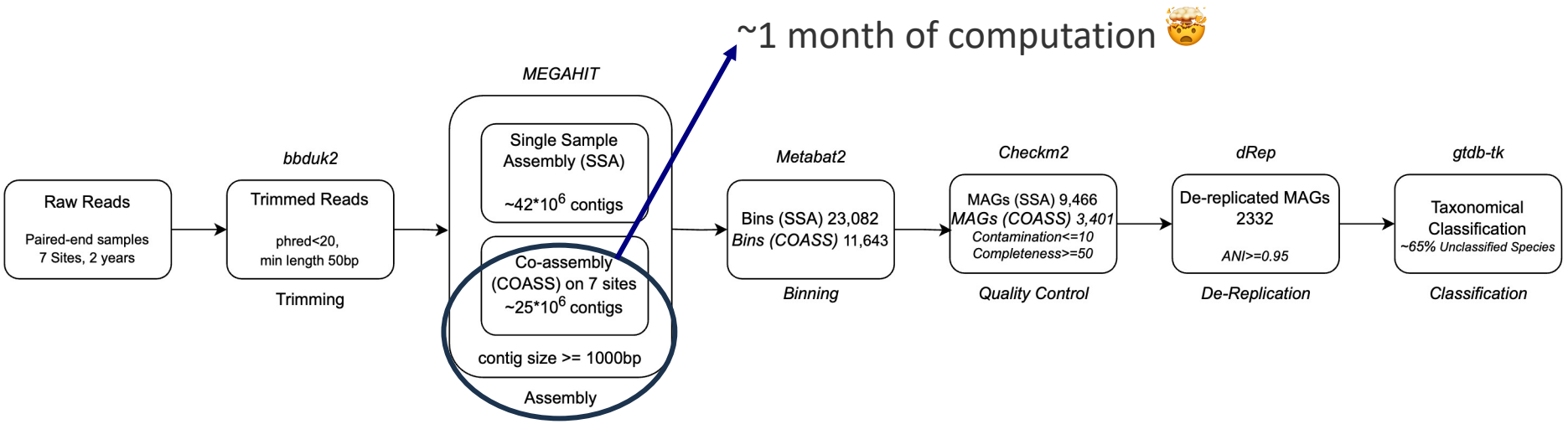
Inferring Microbial Associations pipeline





Developed a pipeline for the analysis of compositional metagenomic data:

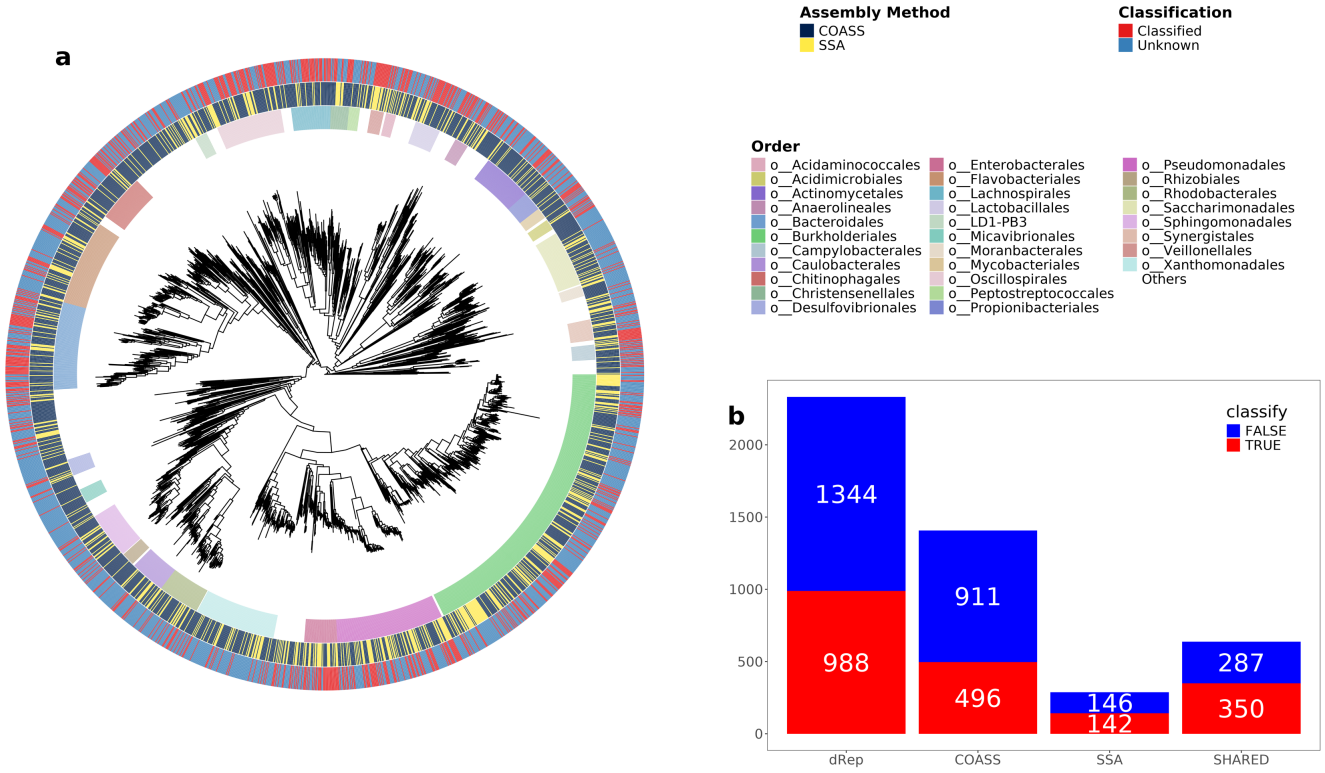
- processing (CLR) & filtering
- calculation of correlation & network construction
- network community identification & visualization
- mgnet R toolbox



Entire pipeline developed and run on the VEO collaborative platform.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Results: 2,332 Reconstructed genomes — Many Still Unknown

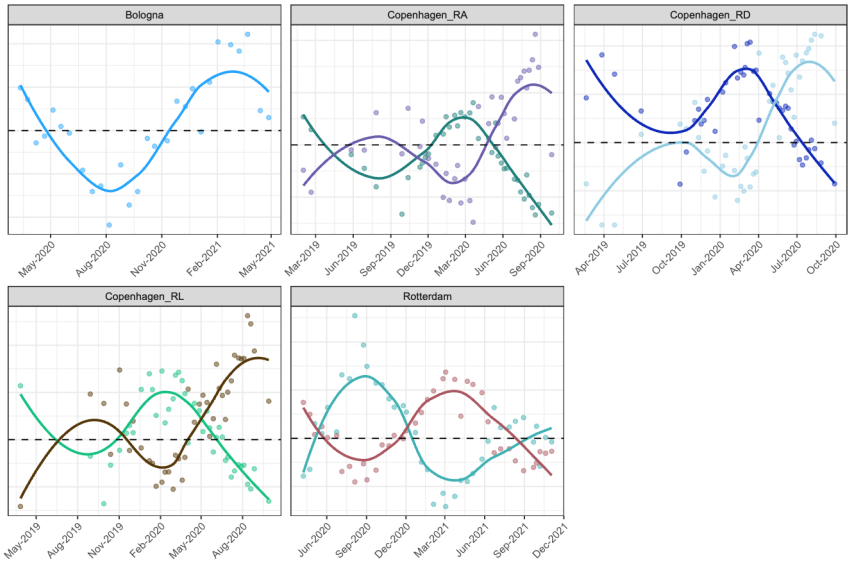


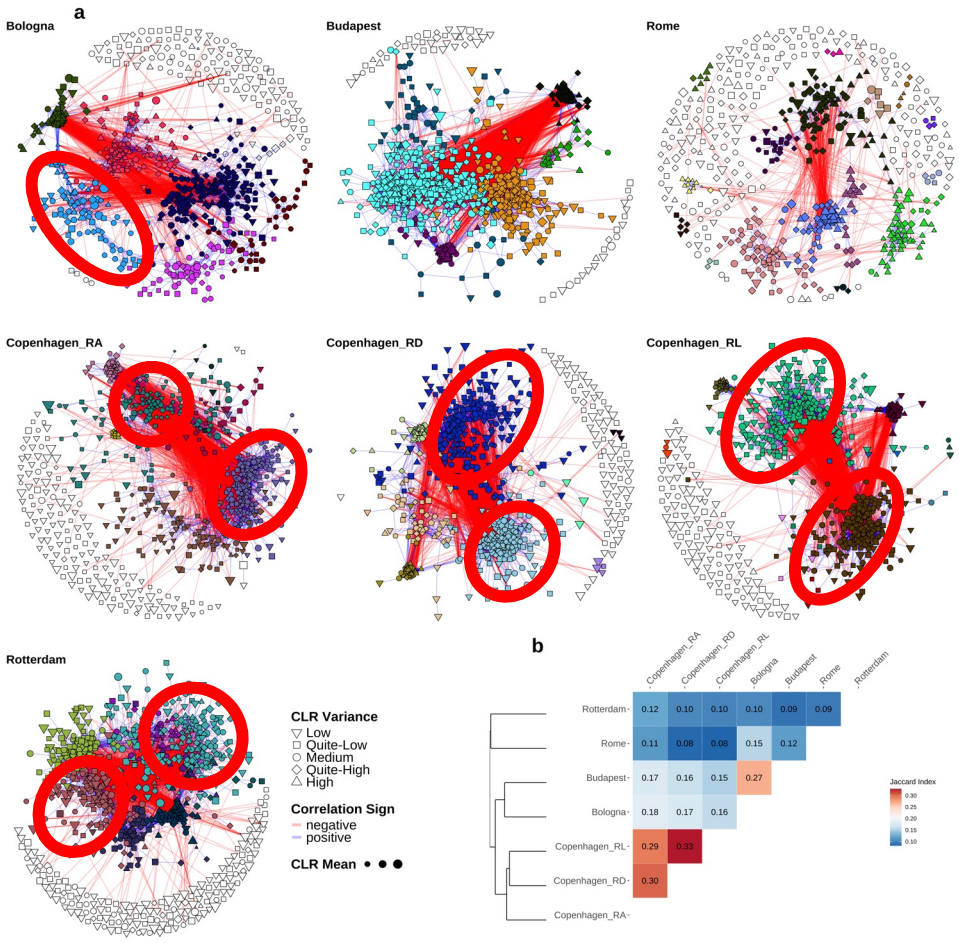
- ~65% of MAGs are unclassified species
- Genomic diversity reflects the complexity of urban wastewater

[Becsei, Fuschi, ..., Remondini, et al. Nature Communications 2024]

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

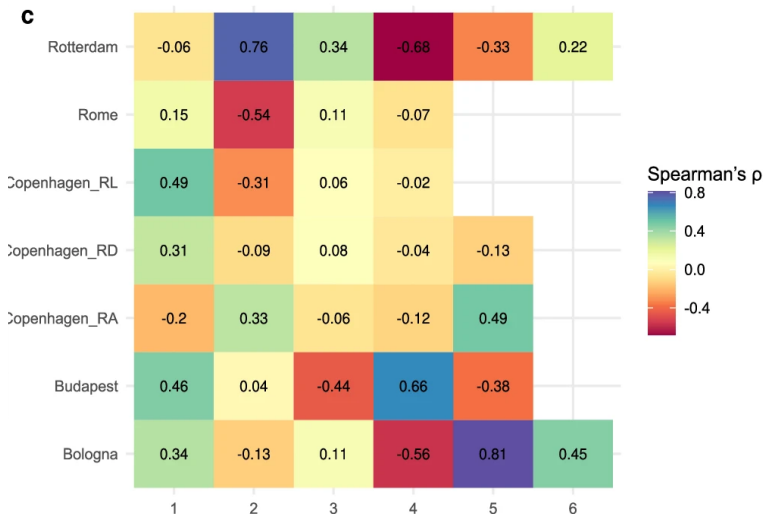
Microbial Community Dynamics

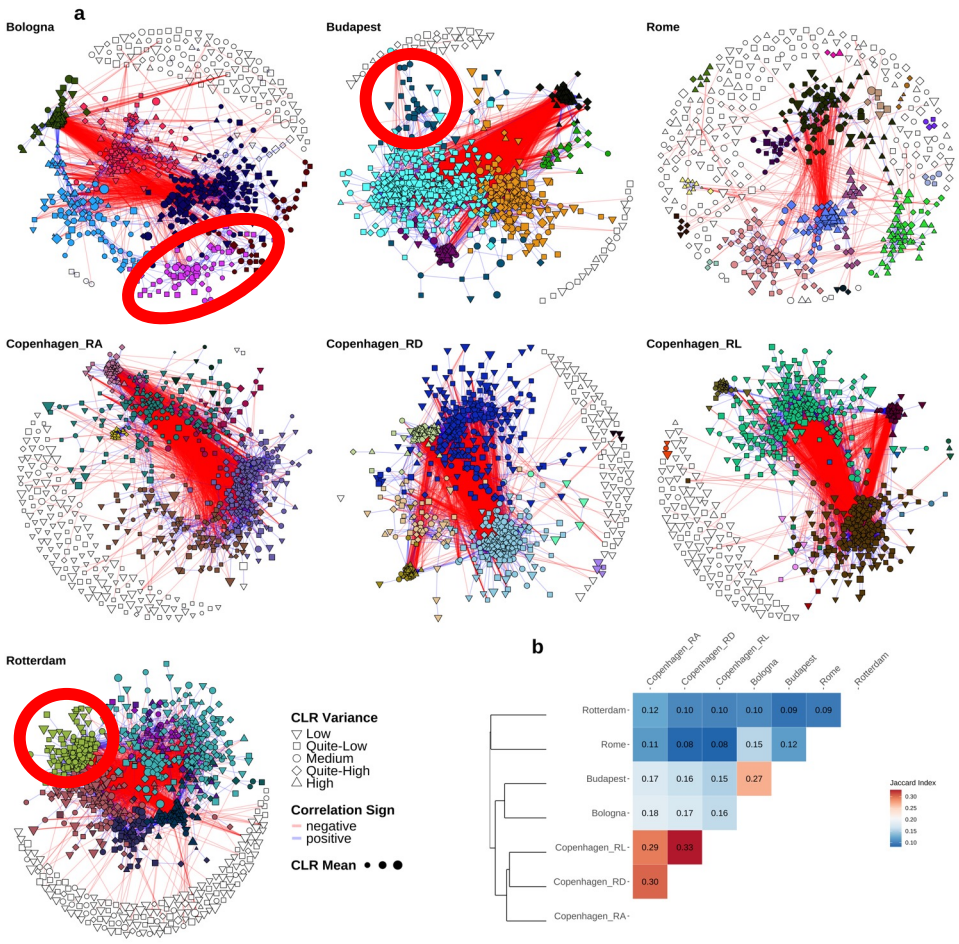




Some communities oscillate in abundance with surprisingly precise 365-day cycles.

Detecting Human-Associated Communities





- **Strong correlations** with *CrAssphage* abundance suggest human origin.

Conclusions

VEO EU project was a great opportunity to develop good research and to apply several physical and analytic methods thanks to:

- project design (clear aims, experiment planning)
- data availability (generated within the consortium & publicly available)
- interdisciplinary collaboration (concepts, terminology, methods)

Acknowledgements

DIFA - UNIBO:
Francesco Durazzi
Alessandro Fuschi
Armando Bazzani
Giulio Colombini
Enrico Lunedei
Alessandra Merlotti

DTU (DK):
Frank Aarestrup
Patrick Munk

EMC (ML):
Marion Koopmans
Ron Fouchier
Miranda De Graaf

ELTE (HU):
Istvan Csabai
Agnes Becsei
David Visontai

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Path Through VEO: From Physics to Metagenomics

🕒 2022

Starting the PhD

- NGS data challenges
- Compositionality
- New metrics & tools


🧪 2023

Bioinformatics & Biology

- Understanding sequencing & biological meaning
- DTU visiting

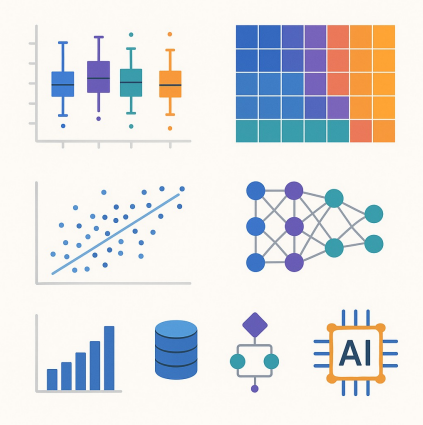



📈 2024

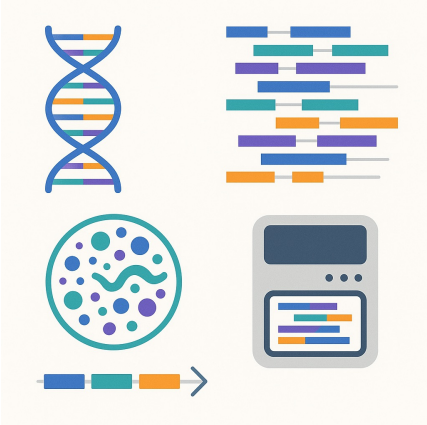
Analysis & Publication

- Statistical & network analysis

Learning Genomics: Not so Easy



Applied physics background 📐
equations, models and coding



Genomics language shock 🤯
What is coverage? Depth? fragment counts? assembly?