Antigen Surveillance: from Evolution to Immune Escape

AI in Pandemic Preparedness and Vaccine Development

Spyros Lytras

25th of March 2026

Outline

1. protein language models (**pLMs**) and **viruses**

2. Unsupervised learning:
   Assessing flu HA site variability with **pLM entropy**

3. Supervised learning:
   Predicting virus fitness with **CoVFit**
   Predicting virus antigenicity with **PLANT**

4. AI for vaccine development

LARGE LANGUAGE MODELS (LLMs)

ChatGPT

Gemini

Training these models on a large corpus of text allows them to *understand* natural language

I like eating sushi 🍣

pronoun verb gerund noun emoji

LARGE LANGUAGE MODELS (LLMs)

Do proteins have a *language*?

You

do proteins have a language? answer in one or two words

Copilot

Using **the web**

Yes, amino acids.

pronoun verb gerund noun emoji

I like eating sushi

M A I S G D D C

start codon

RdRp motif

I like eating sushi	M A I S G D D C	
I **sushi** eating sushi	M A I S G **A** **A** C	**grammar**
I **don't** like eating sushi	M **V** A I S G D D C	**meaning**
I like eating **croissants**	M A **G** S G D D **A**	**context**

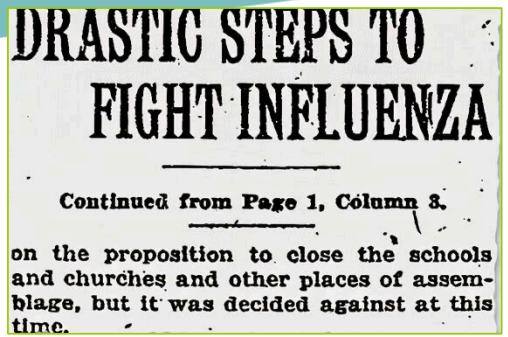
Learning the language of viral evolution and escape

BRIAN HIE, ELLEN D. ZHONG, BONNIE BERGER, AND BRYAN BRYSON Authors Info & Affiliations

SCIENCE • 15 Jan 2021 • Vol 371, Issue 6526 • pp. 284-288 • DOI: 10.1126/science.abd7331

Grammar = viability/fitness

Meaning = Antigenic variation



DRASTIC STEPS TO FIGHT INFLUENZA

Continued from Page 1, Column 8.

on the proposition to close the schools and churches and other places of assemblage, but it was decided against at this time.

The Brooklyn Eagle 10 cents ★★★★★

A Pulitzer Prize-Winning Newspaper JA 2-3100 Tuesday, February 19, 1963

Asian Flu Here Called Epidemic

By GENE RUFFINI

Health

First presumptive positive case of H5 avian influenza detected in B.C.

ecdc EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

Detection of avian flu antibodies in Dutch dairy cow: ECDC risk assessment remains unchanged

LA GRIPPE SÉVIT dans notre RÉGION...

Mais il n'y a pas lieu de s'affoler

2 jours

A Grisy-Suisnes

Un bambin

SWINE FLU SPREADS!

Feds fear virus could turn deadly in U.S.

Mayor says don't panic as cases confirmed

EVERYTHING YOU NEED TO KNOW — SEE PAGES 4-6

A year later, cow flu origins are an unsettling puzzle

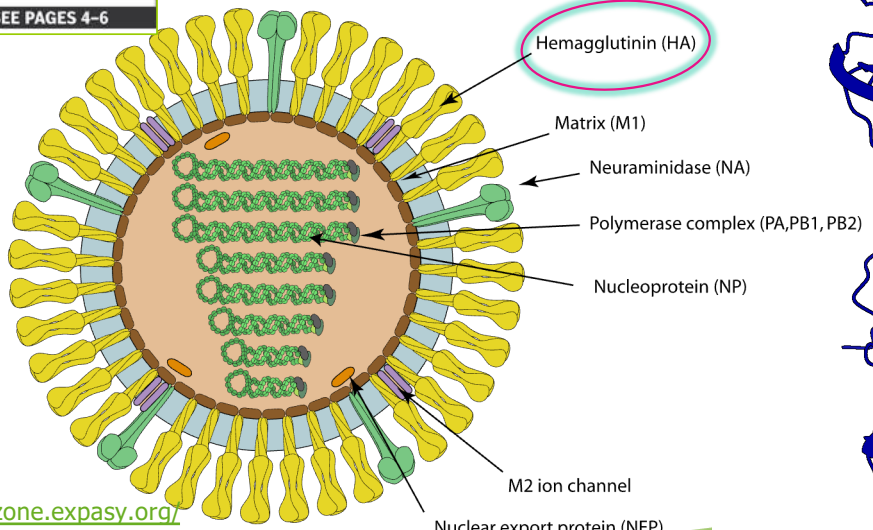
It's still unclear how H5N1 virus jumped into U.S. cattle—and why it keeps doing so

25 MAR 2025 • 5:35 PM ET • BY KAI KUPFERSCHMIDT

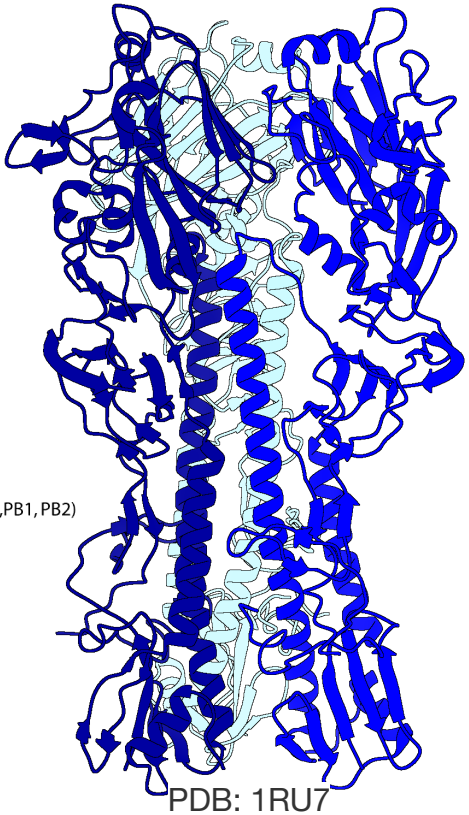
Bird flu H5N1 precautions ramp up as thousands of dead birds wash up along east coastline

By Romy Gilbert | ABC Illawarra | Birds

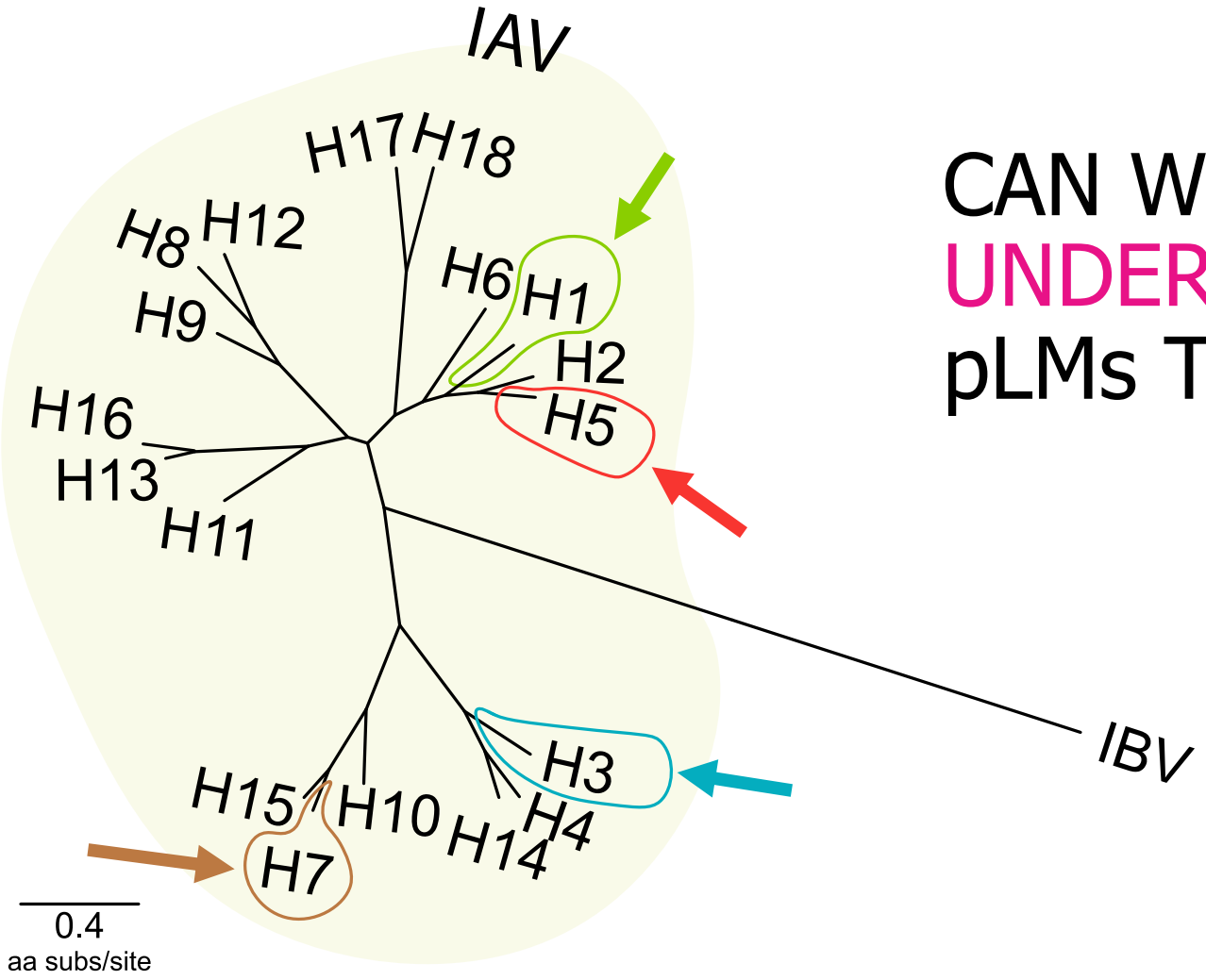
20h ago



Viralzone
https://viralzone.expasy.org/



PDB: 1RU7

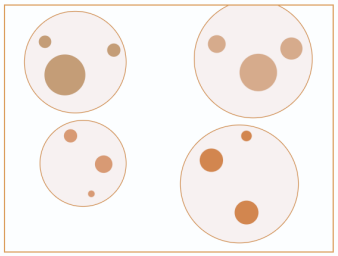


CAN WE FOCUS THE UNDERSTANDING OF pLMs TO THE FLU HA?

ESM-2

Training dataset (corpus of protein sequences):

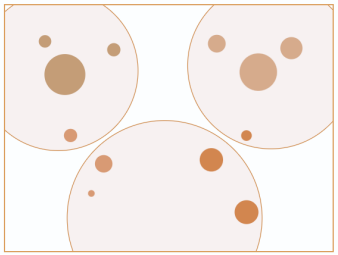
UniRef100



UniRef100 contains all UniProt Knowledgebase records plus selected UniParc records. In UniRef100, all identical sequences and subfragments with 11 or more residues are placed into a single record.

Identity percentage	Number of UniRef clusters	Download
100%	453,950,711 clusters	README FTP

UniRef90



UniRef90 is generated by clustered UniRef100 sequences with 11 or more residues, such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the seed sequence.

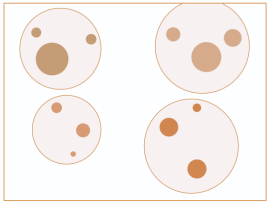
Identity percentage	Number of UniRef clusters	Download
90%	204,806,910 clusters	README FTP

Overall *understanding* of a range of diverse protein sequences representing all the tree of life **(but probably not many virus sequences)**

FINE-TUNE THE ORIGINAL MODEL TO THE FLU HA DIVERSITY

(unsupervised learning)

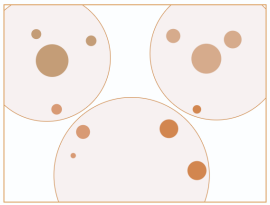
UniRef100



UniRef100 contains all UniProt Knowledgebase records plus selected UniParc records. In UniRef100, all identical sequences and subfragments with 11 or more residues are placed into a single record.

Identity percentage	Number of UniRef clusters	Download
100%	453,950,711 clusters	README FTP

UniRef90



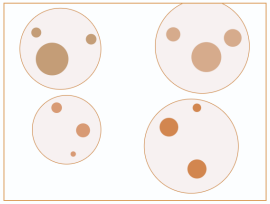
UniRef90 is generated by clustered UniRef100 sequences with 11 or more residues, such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the seed sequence.

Identity percentage	Number of UniRef clusters	Download
90%	204,806,910 clusters	README FTP

FINE-TUNE THE ORIGINAL MODEL TO THE FLU HA DIVERSITY

(unsupervised learning)

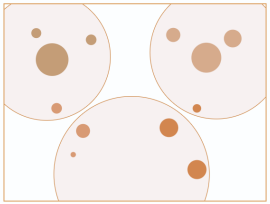
UniRef100



UniRef100 contains all UniProt Knowledgebase records plus selected UniParc records. In UniRef100, all identical sequences and subfragments with 11 or more residues are placed into a single record.

Identity percentage	Number of UniRef clusters	Download
100%	453,950,711 clusters	README FTP

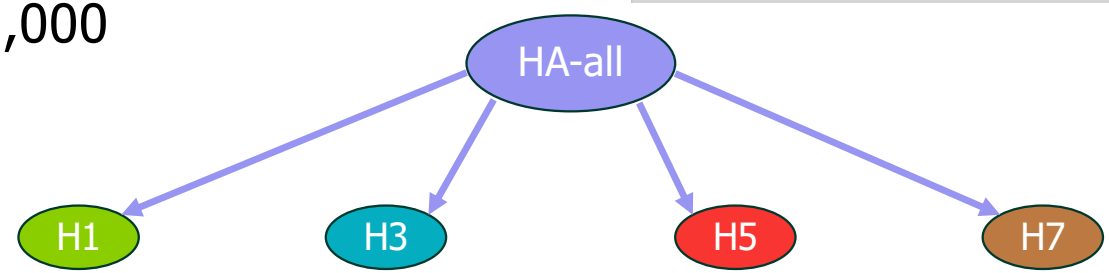
UniRef90



UniRef90 is generated by clustered UniRef100 sequences with 11 or more residues, such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the seed sequence.

Identity percentage	Number of UniRef clusters	Download
90%	204,806,910 clusters	README FTP

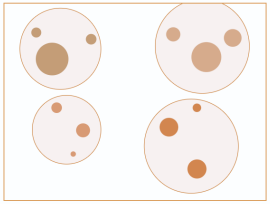
All HA sequences in NCBIflu clustered by 99% similarity ~10,000



FINE-TUNE THE ORIGINAL MODEL TO THE FLU HA DIVERSITY

(unsupervised learning)

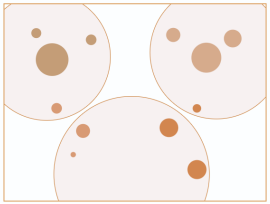
UniRef100



UniRef100 contains all UniProt Knowledgebase records plus selected UniParc records. In UniRef100, all identical sequences and subfragments with 11 or more residues are placed into a single record.

Identity percentage	Number of UniRef clusters	Download
100%	453,950,711 clusters	README FTP

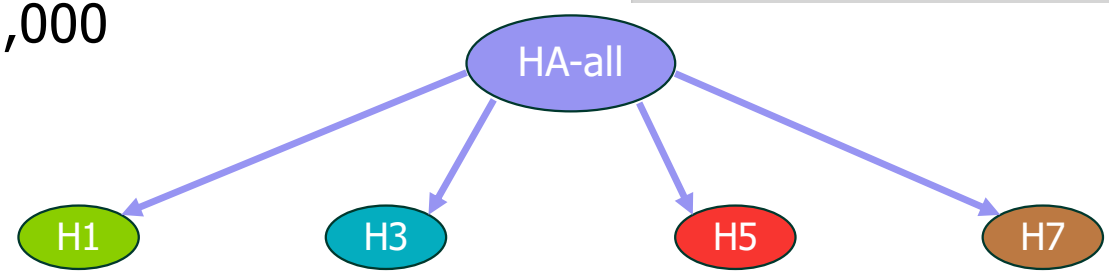
UniRef90



UniRef90 is generated by clustered UniRef100 sequences with 11 or more residues, such that each cluster is composed of sequences that have at least 90% sequence identity to and 80% overlap with the seed sequence.

Identity percentage	Number of UniRef clusters	Download
90%	204,806,910 clusters	README FTP

All HA sequences in NCBIflu clustered by 99% similarity ~10,000



ESM-2
HA-all

ESM-2
H1

ESM-2
H3

ESM-2
H5

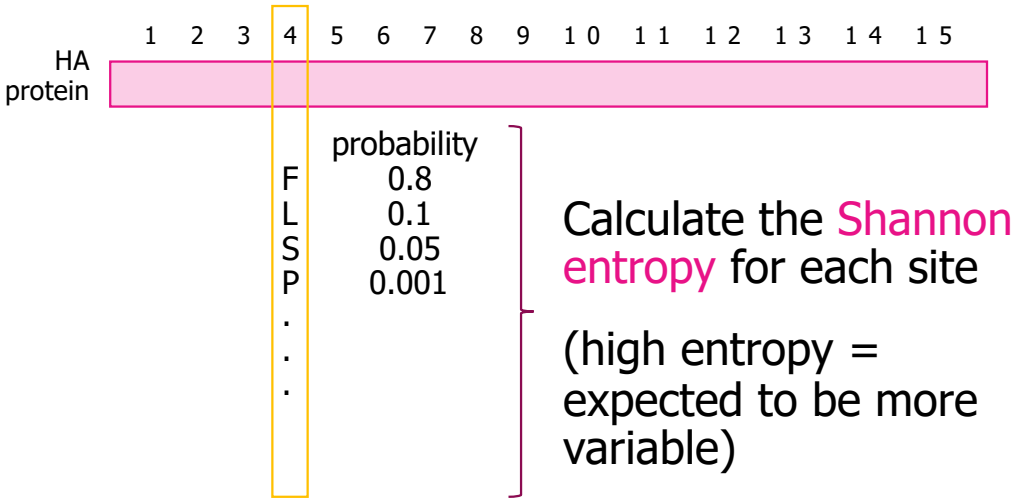
ESM-2
H7

WE HAVE A FLU-SPECIFIC pLM!

WHAT DO WE USE IT FOR? 🤔

The model can infer a probability of any amino acid in a given sequence context.

pLM entropy

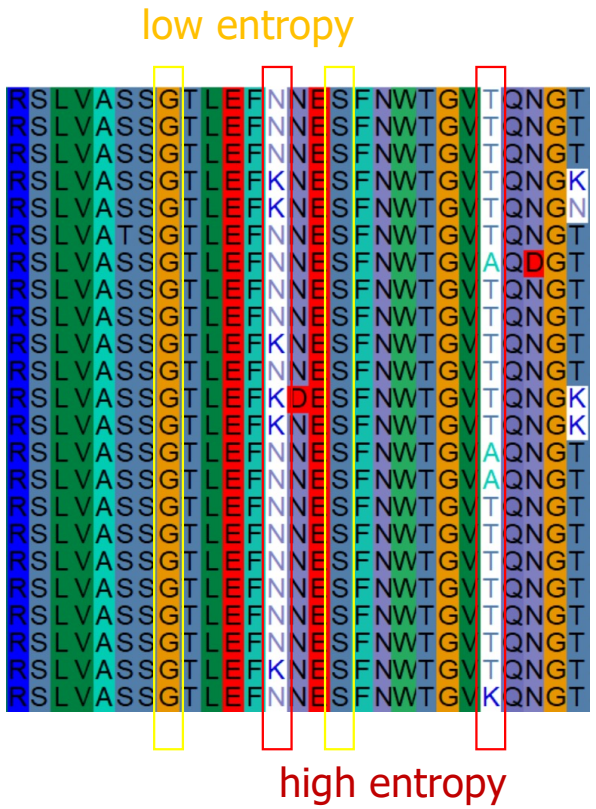


Calculate the Shannon's entropy for each site

(high entropy = expected to be more variable)

Specific for a single (unaligned) sequence

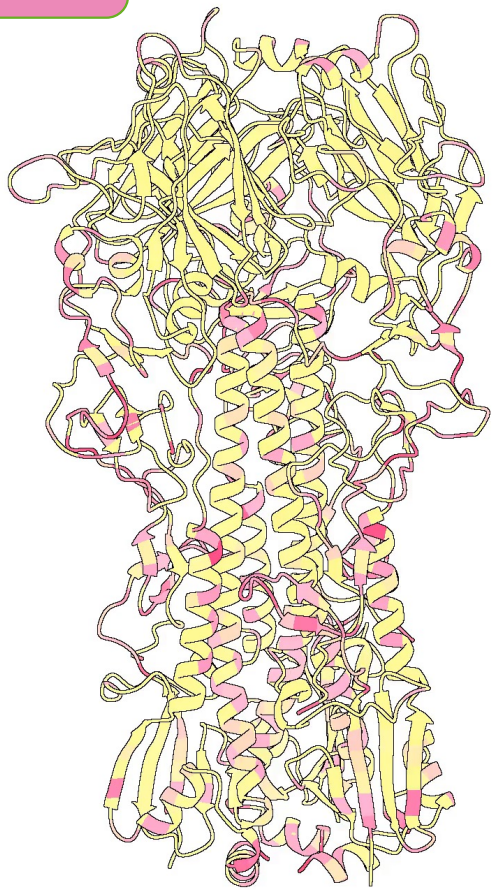
Alignment entropy



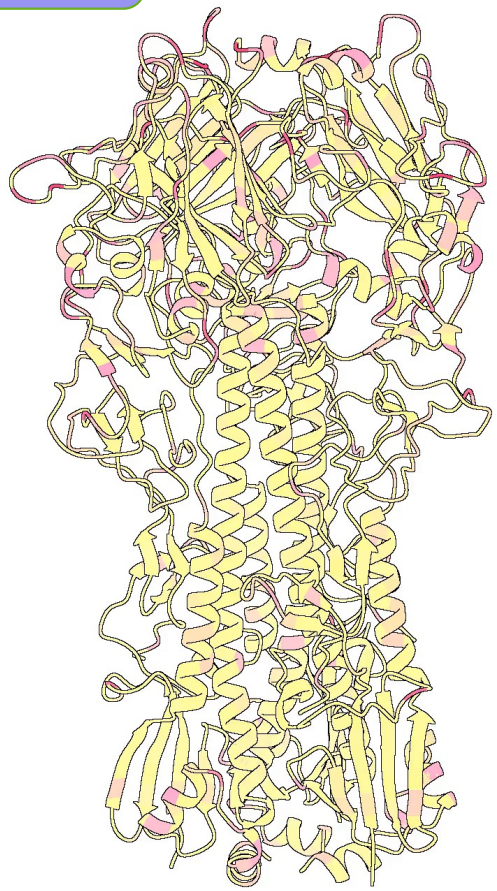
Specific for a multiple sequence alignment

Compare average **pLM entropy** to alignment entropy

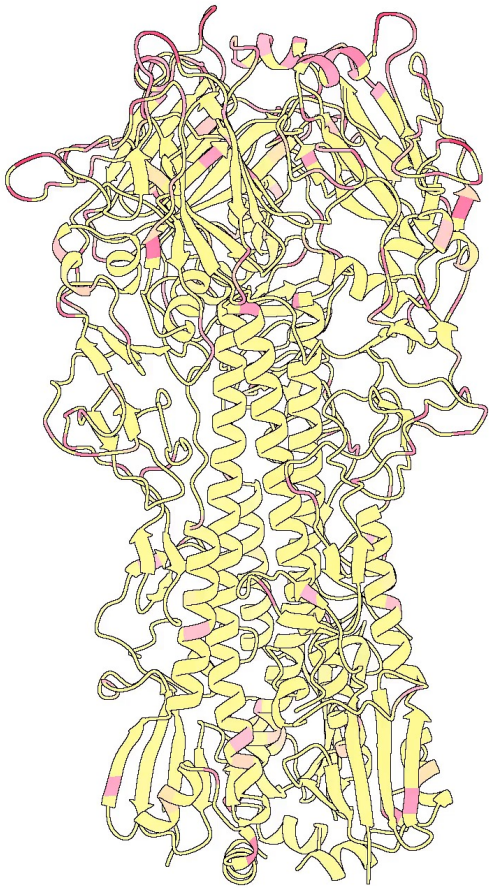
ESM-2



ESM-2
HA-all



H3
alignment

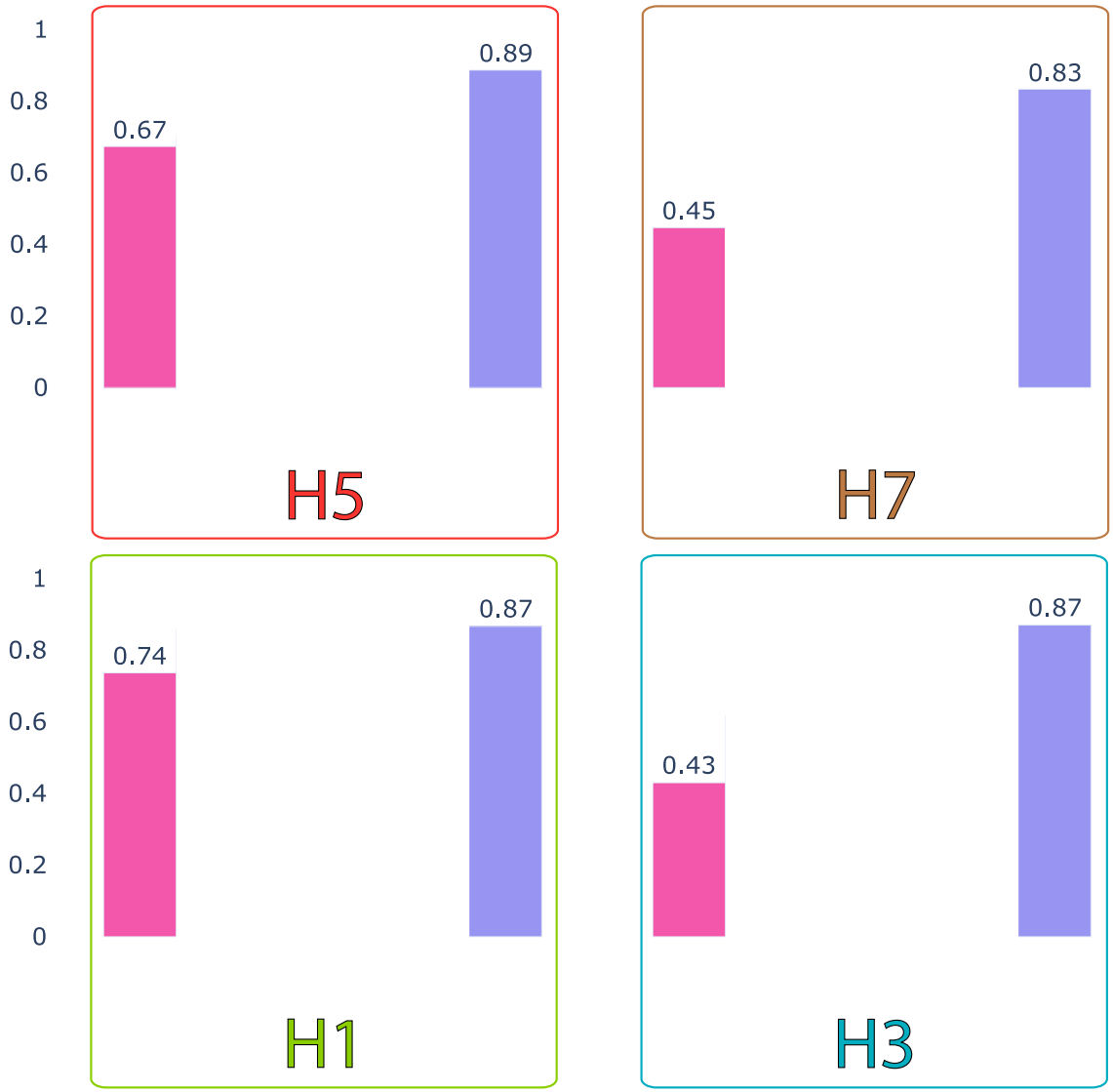


Average entropy score for multiple H3 HAs

Spearman correlation between per-site **pLM entropy** and **alignment entropy**

pLM entropy correlates with alignment entropy

ESM-2

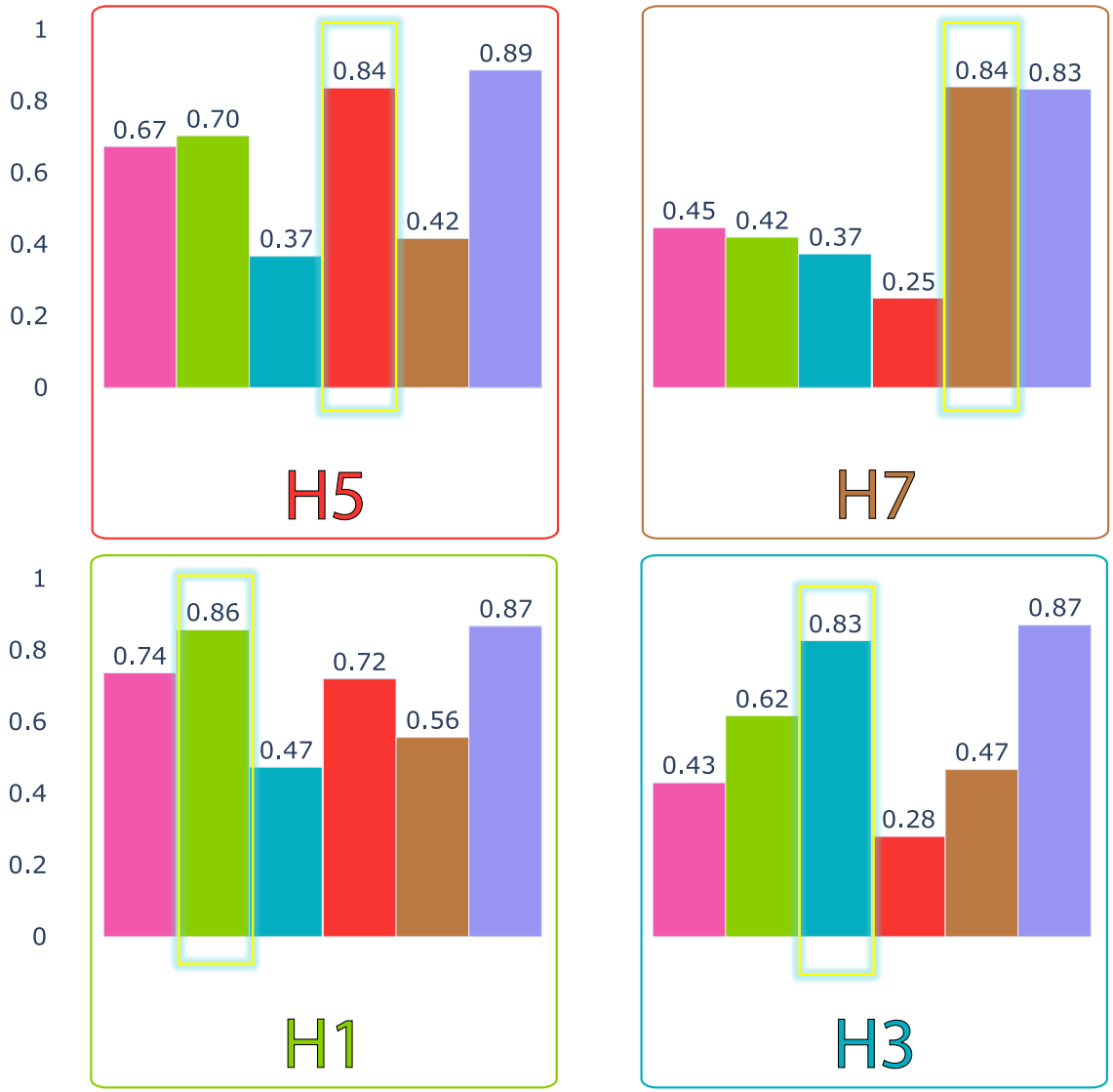


Spearman correlation between per-site **pLM entropy** and **alignment entropy**

pLM entropy correlates with alignment entropy

Fine-tuned model has substantially better performance

ESM-2 | ESM-2 HA-all

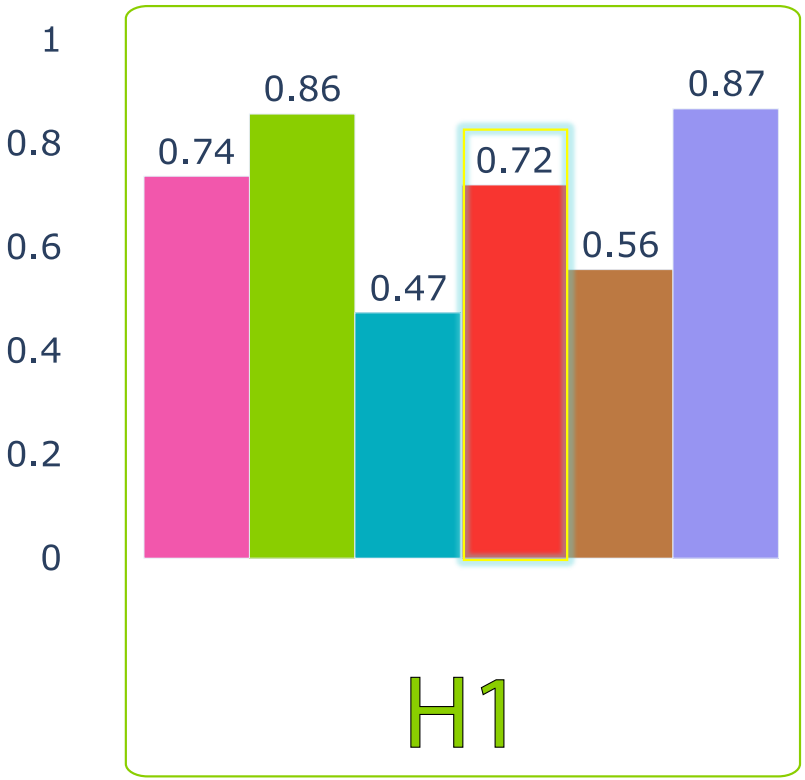


Spearman correlation between per-site **pLM entropy** and **alignment entropy**

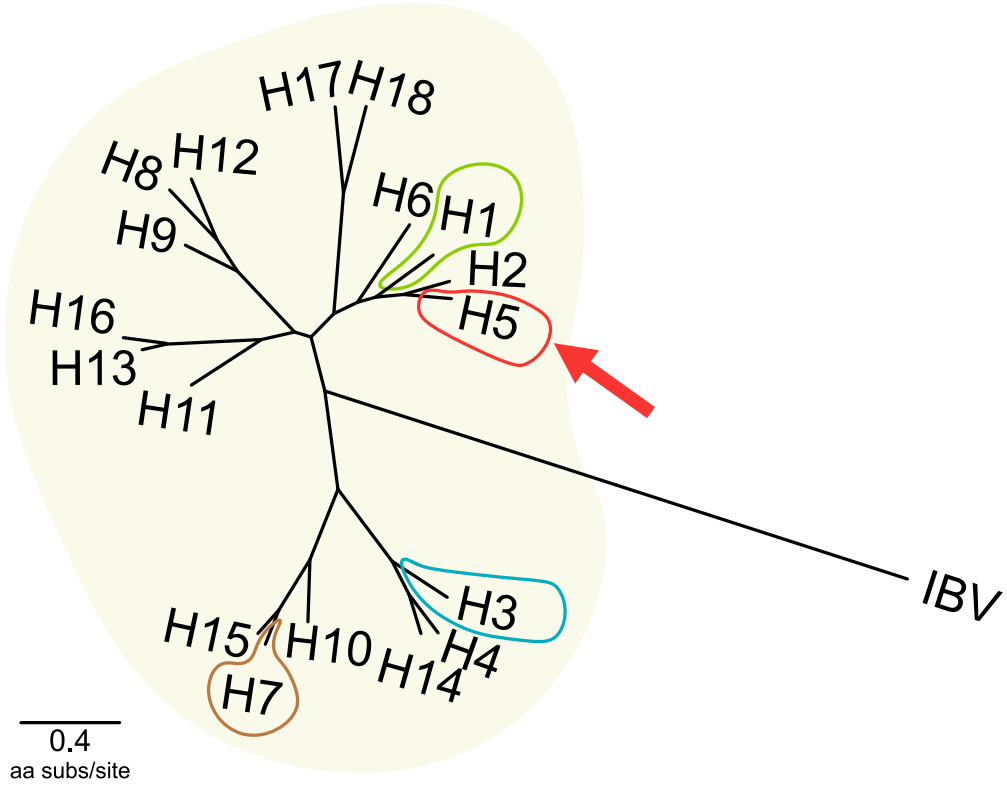
How do different training datasets compare

Fine-tuning on one serotype improves performance for that serotype (~only)

How do different training datasets compare

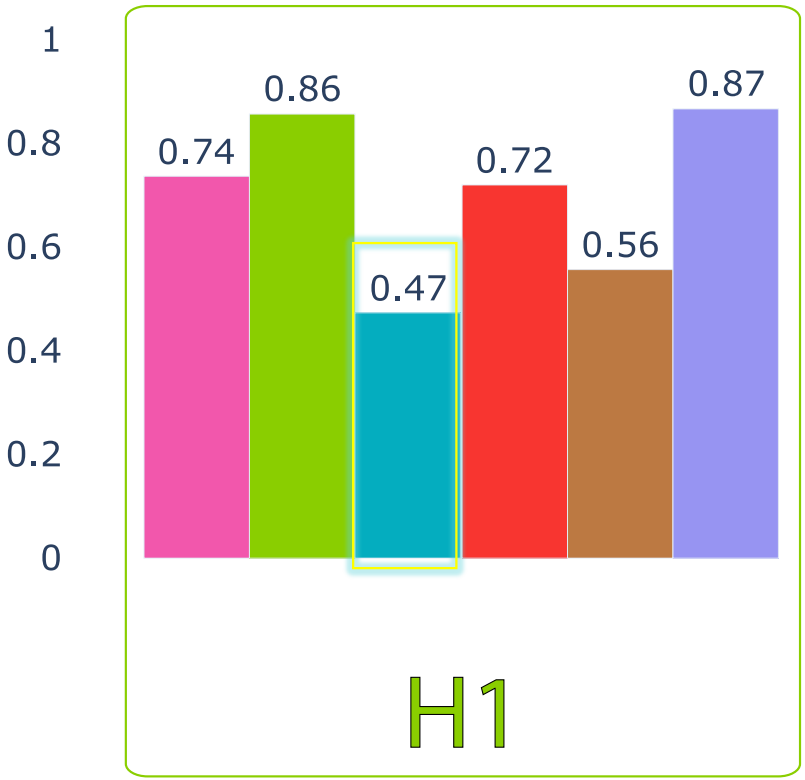


Spearman correlation between per-site **pLM entropy** and **alignment entropy**



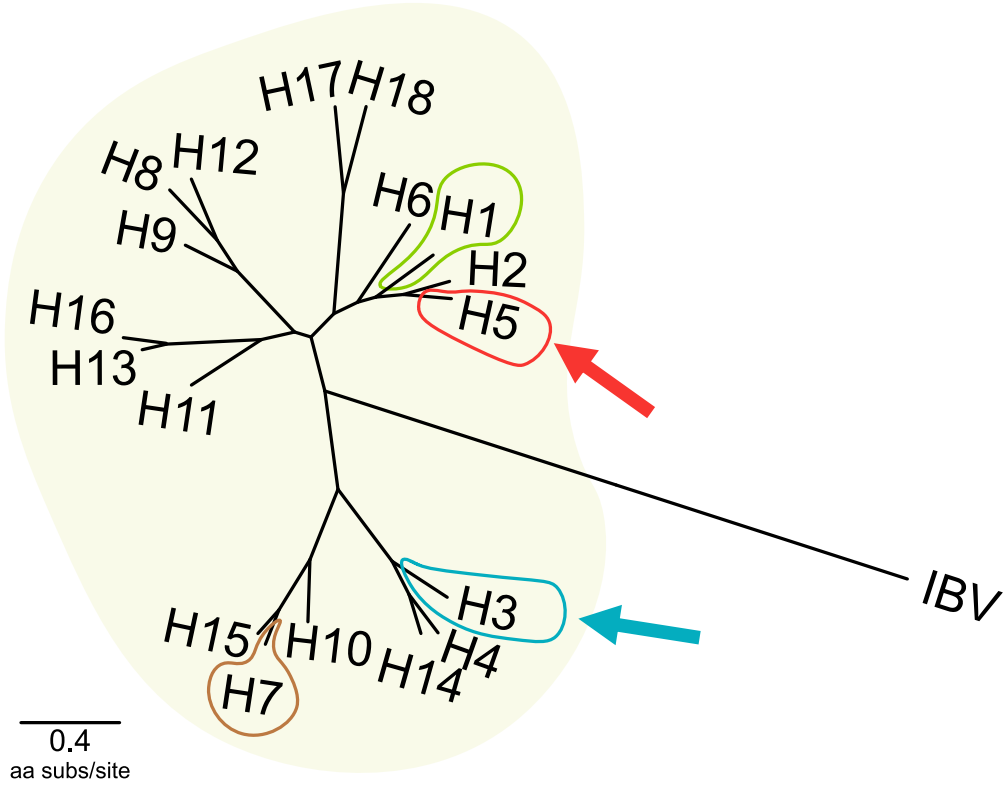
ESM-2 | ESM-2 HA-all | ESM-2 H1 | ESM-2 H3 | ESM-2 H5 | ESM-2 H7

FINE-TUNING ON RELATED BUT **DISTANT** DIVERSITY **WORSENS** PERFORMANCE



Spearman correlation between per-site **pLM entropy** and **alignment entropy**

How do different training datasets compare

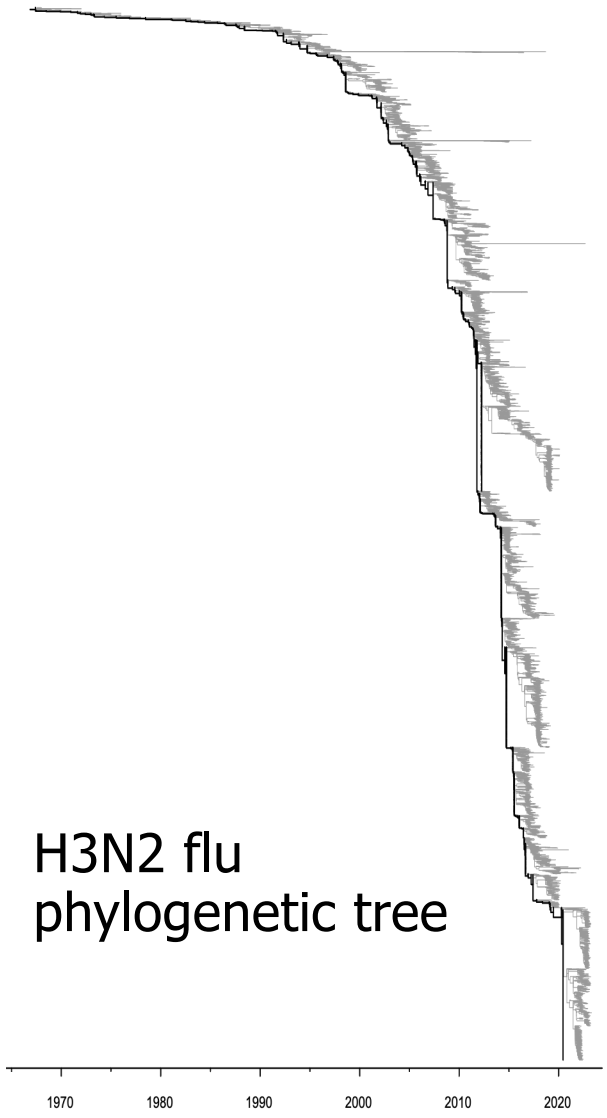


ESM-2 | ESM-2 HA-all | ESM-2 H1 | ESM-2 H3 | ESM-2 H5 | ESM-2 H7

pLM entropy correlates with **alignment entropy** and performance improves when the pLM is **fine-tuned** on the protein group of interest

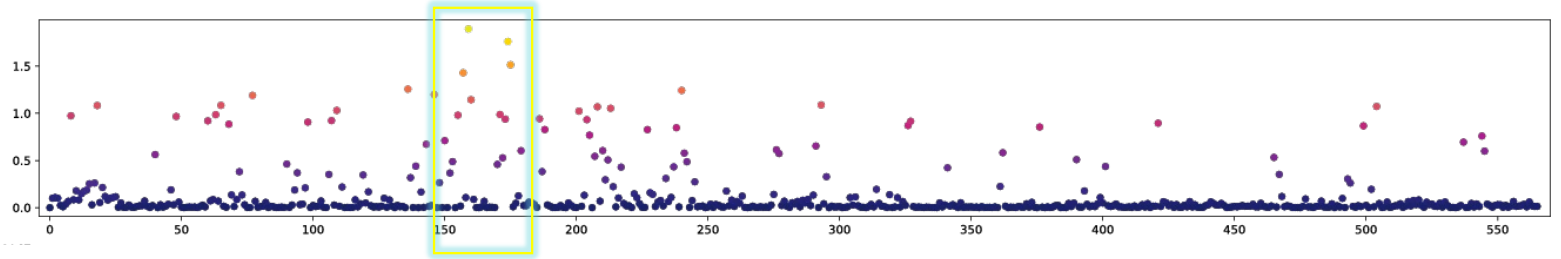
pLM entropy is **sequence-specific**:

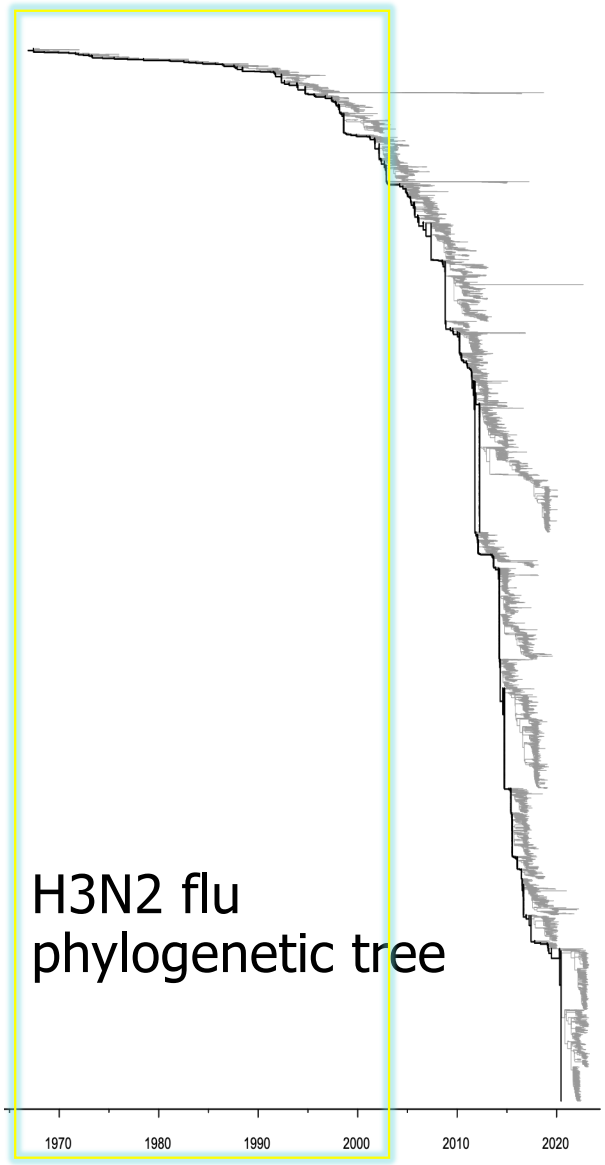
no alignment
single sequence only



H3N2 flu
phylogenetic tree

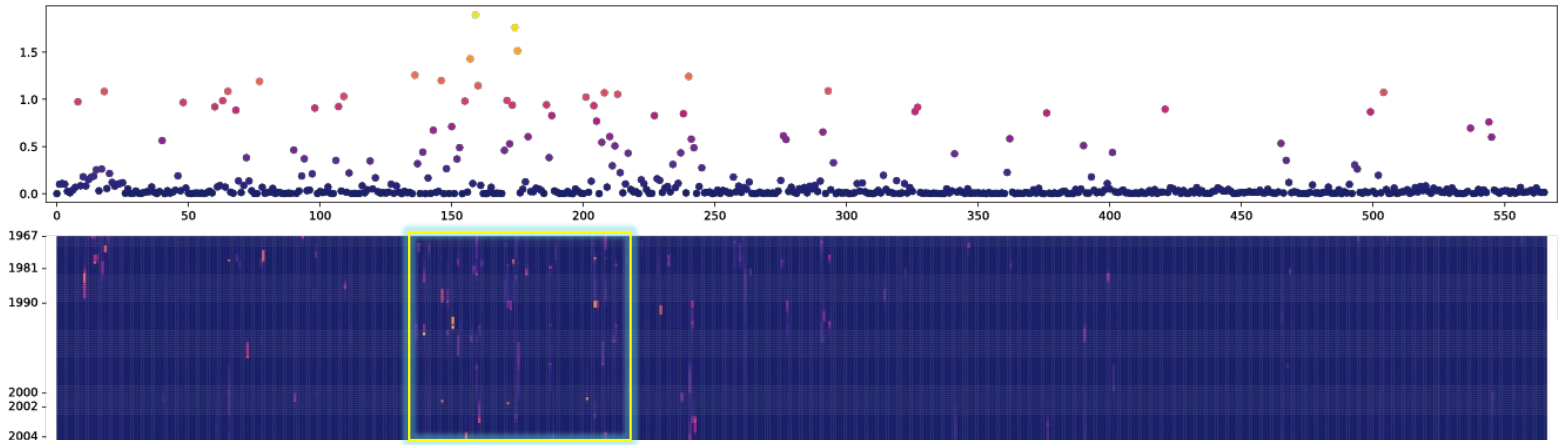
alignment entropy



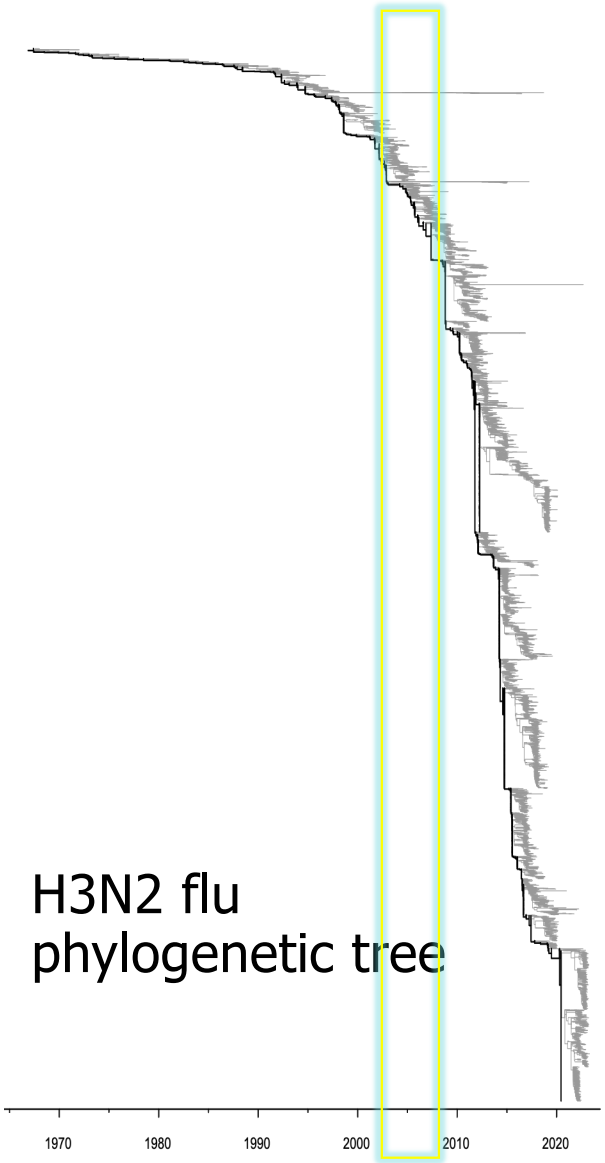


H3N2 flu
phylogenetic tree

alignment entropy

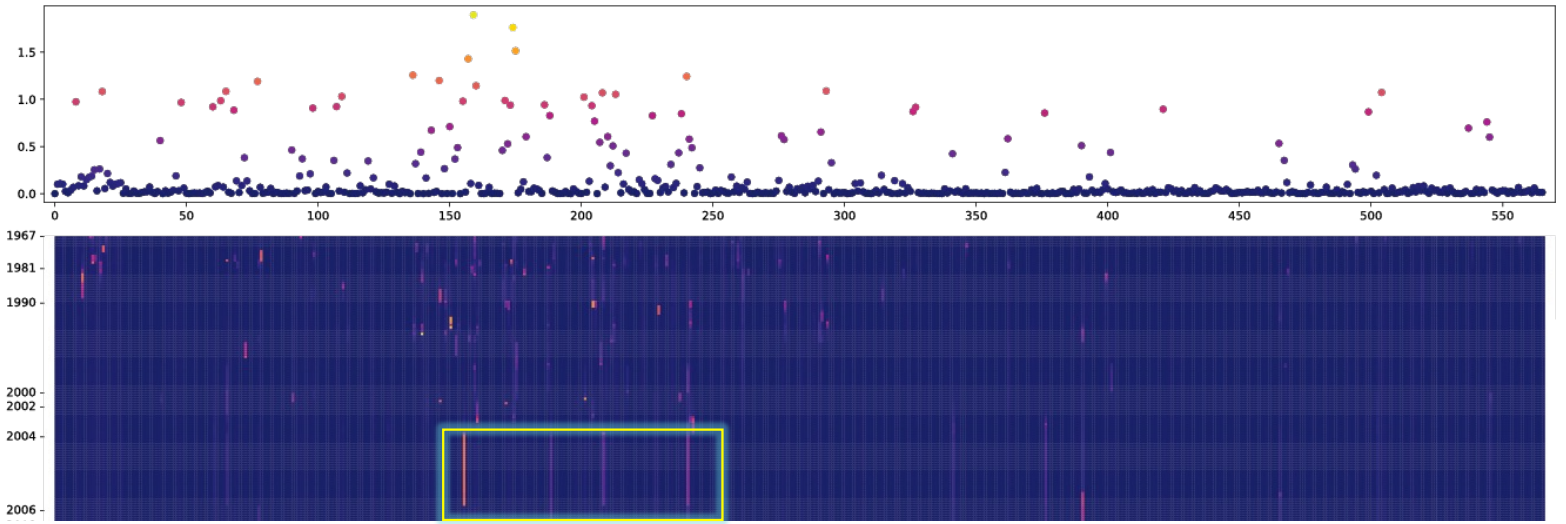


pLM entropy

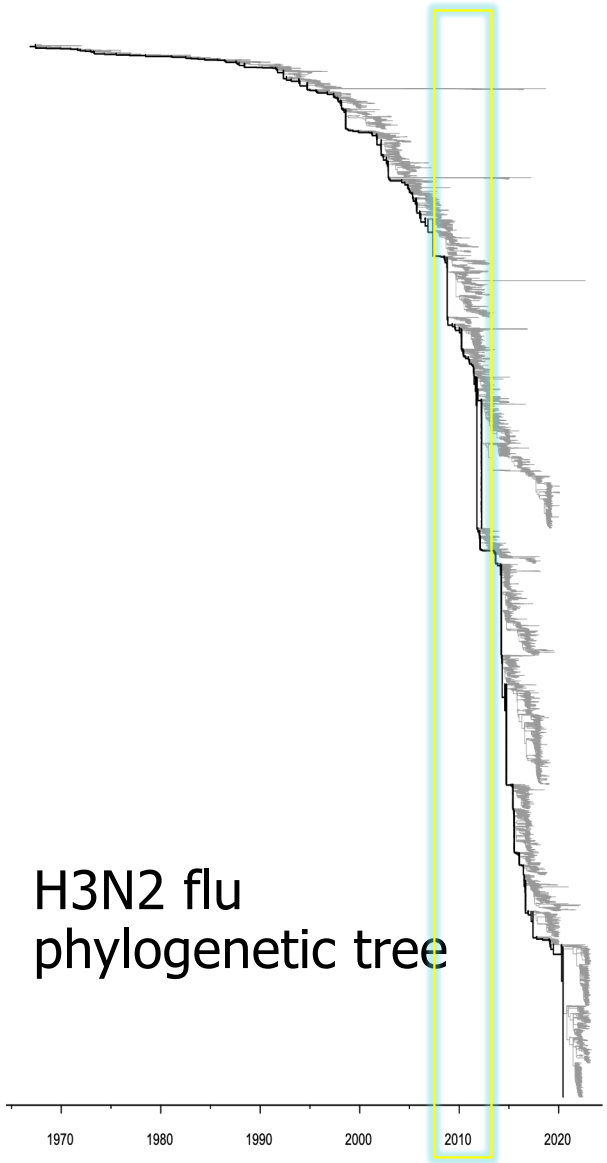


H3N2 flu
phylogenetic tree

alignment entropy

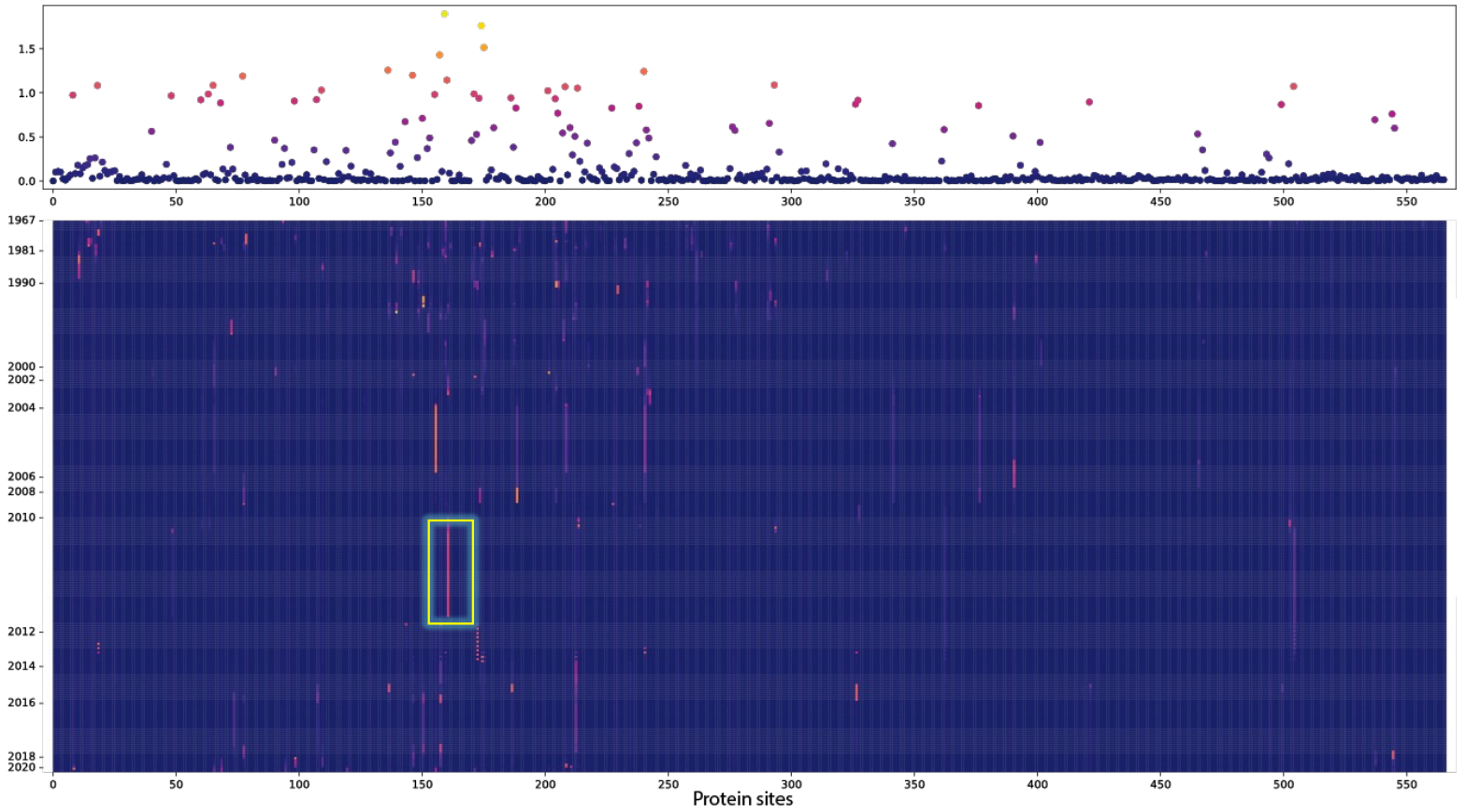


pLM entropy

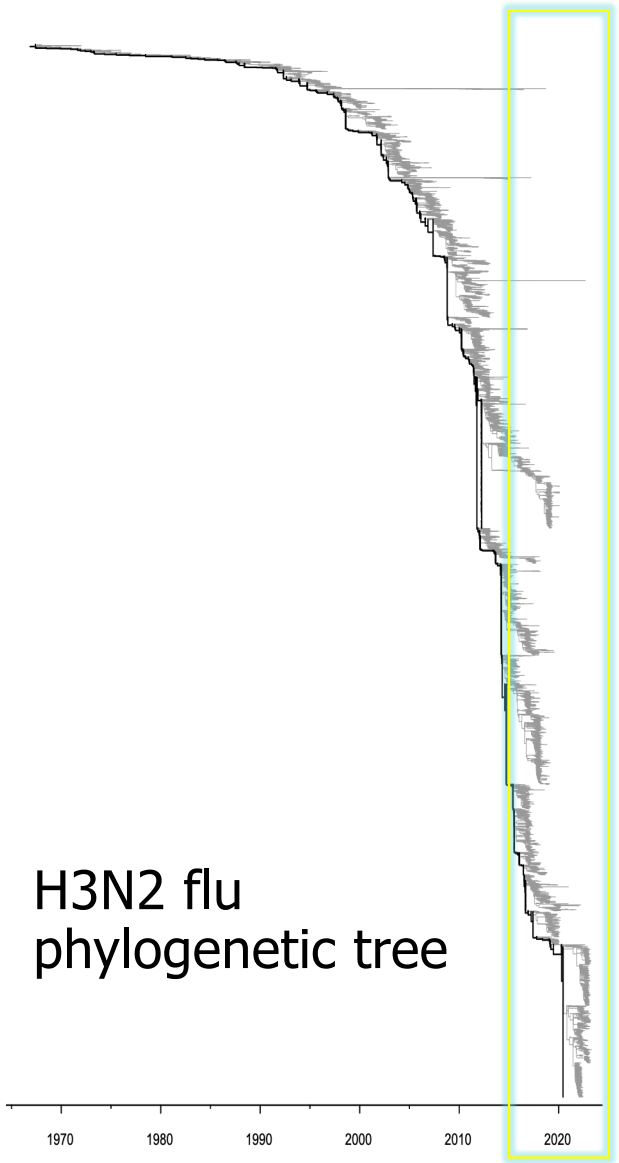


H3N2 flu
phylogenetic tree

alignment entropy

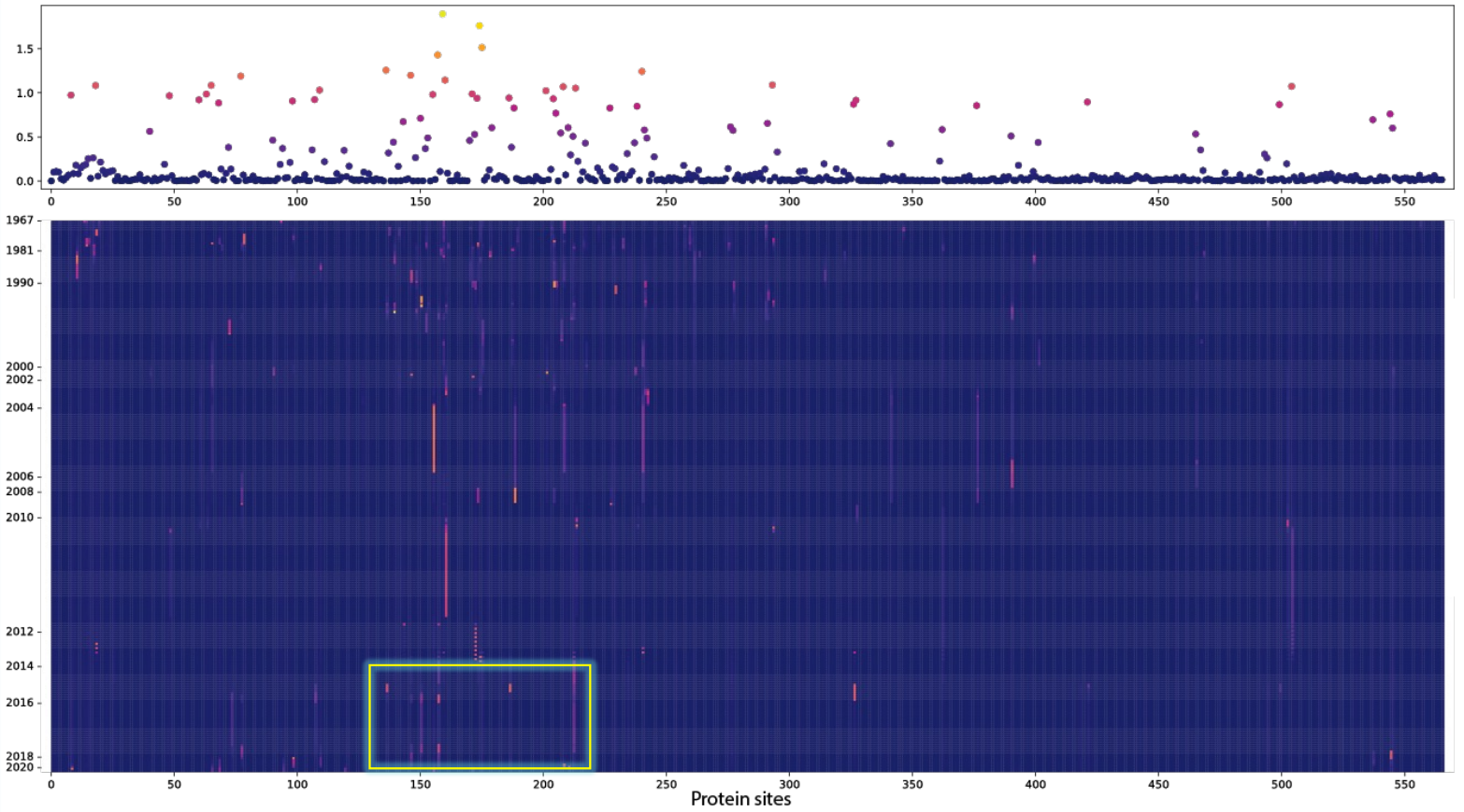


pLM entropy



H3N2 flu
phylogenetic tree

alignment entropy



pLM entropy

pLM entropy can predict mutable sites in a sequence

Fine-tuning on a specific task

(supervised learning)

Training sequence dataset	Associated values
M A I S G D D C	0.1
M A I S G **A** **A** C	0.5
M **V** A I S G D D C	0.3
M A **G** S G D D **A**	0.8

Link genotype to phenotype

Jumpei Ito

VIRUS INFORMATICS

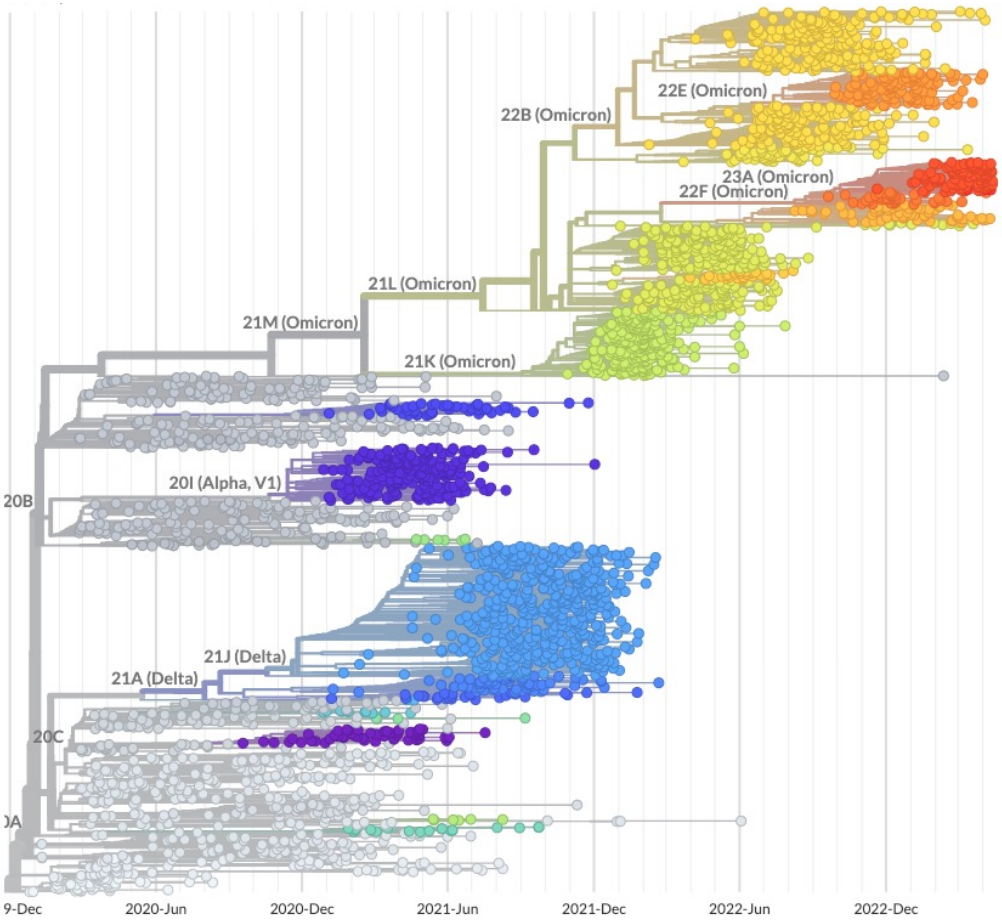
ITO Laboratory

RIMD

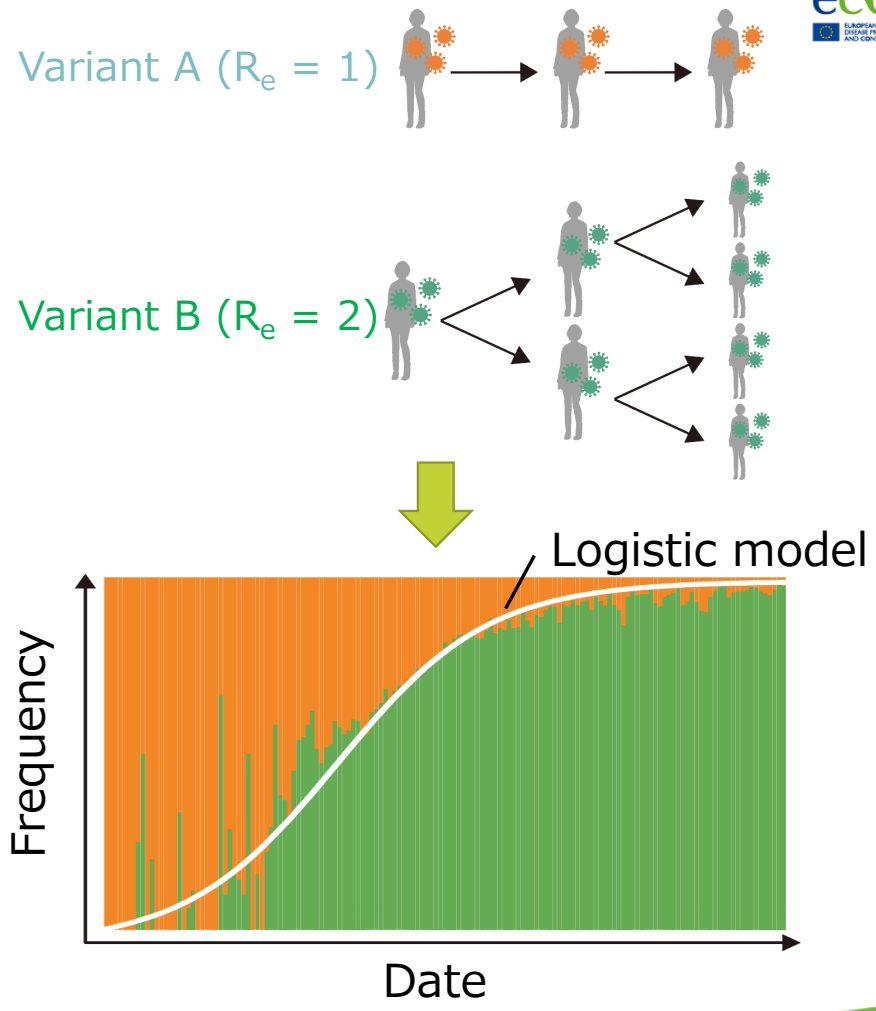
Research Institute for Microbial Diseases

大阪大学微生物病研究所

Estimating the reproduction number (R_e) of SARS-CoV-2 variants based on genomic data



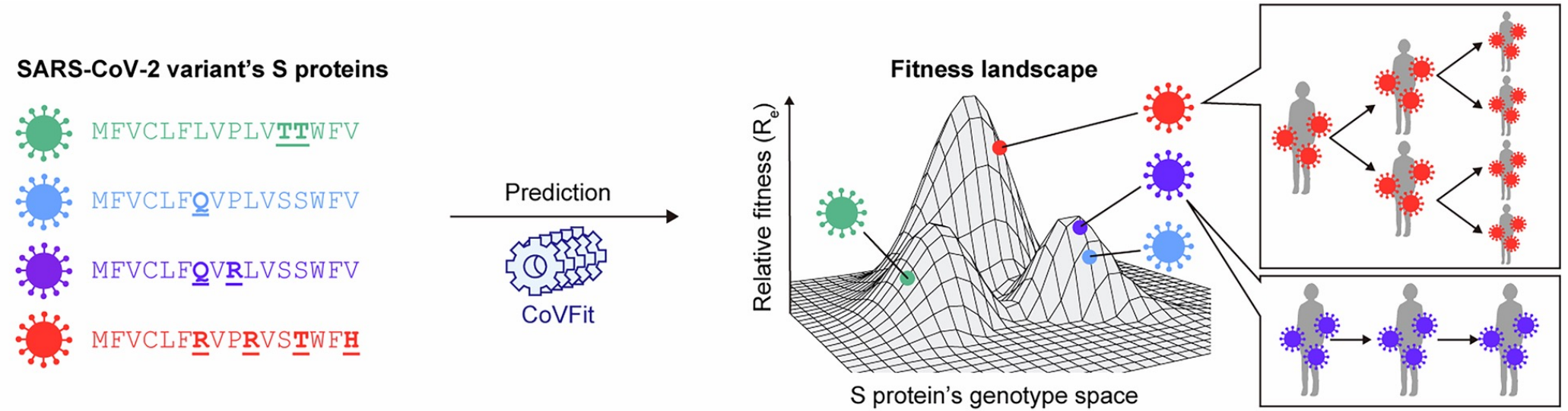
Nextstrain (https://nextstrain.org/)



A protein language model for exploring viral fitness landscapes

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium & Kei Sato

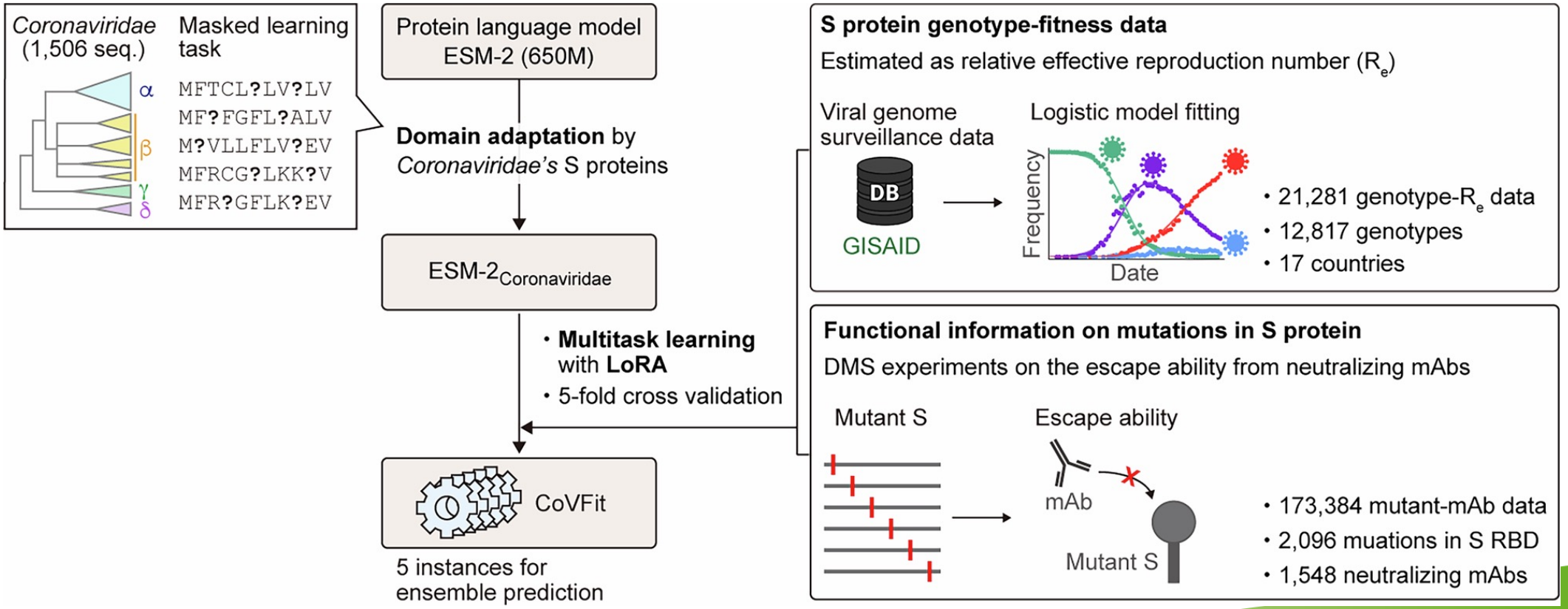
Nature Communications **16**, Article number: 4236 (2025) | Cite this article

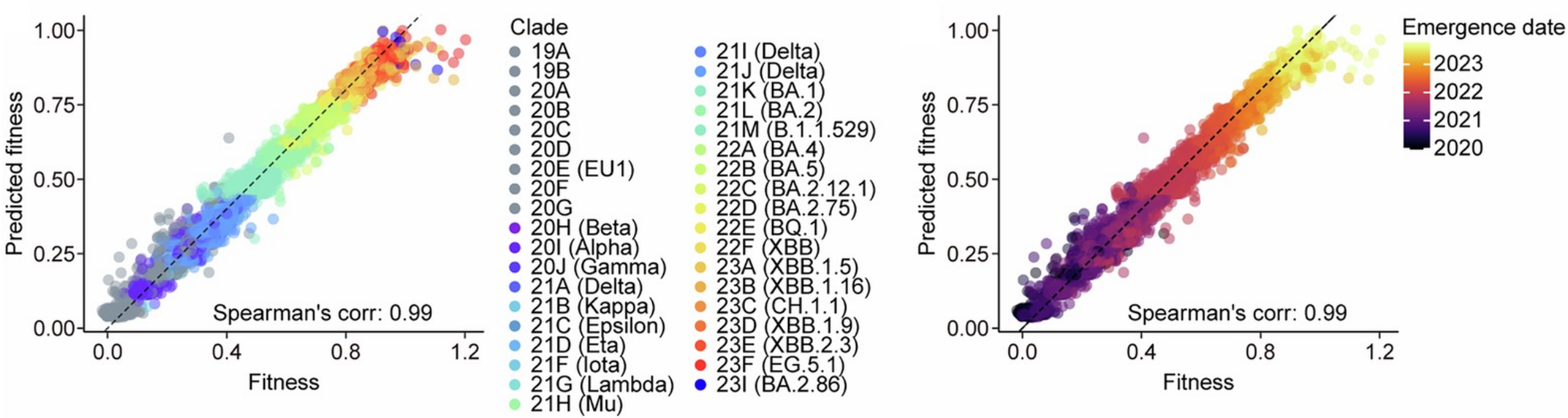


A protein language model for exploring viral fitness landscapes

Jumpei Ito, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium & Kei Sato

Nature Communications **16**, Article number: 4236 (2025) | Cite this article




Predicted fitness
1.00
0.75
0.50
0.25
0.00
0.0
0.4
0.8
1.2
Fitness
Spearman's corr: 0.99
Clade
19A
19B
20A
20B
20C
20D
20E (EU1)
20F
20G
20H (Beta)
20I (Alpha)
20J (Gamma)
21A (Delta)
21B (Kappa)
21C (Epsilon)
21D (Eta)
21F (Iota)
21G (Lambda)
21H (Mu)
21I (Delta)
21J (Delta)
21K (BA.1)
21L (BA.2)
21M (B.1.1.529)
22A (BA.4)
22B (BA.5)
22C (BA.2.12.1)
22D (BA.2.75)
22E (BQ.1)
22F (XBB)
23A (XBB.1.5)
23B (XBB.1.16)
23C (CH.1.1)
23D (XBB.1.9)
23E (XBB.2.3)
23F (EG.5.1)
23I (BA.2.86)
Predicted fitness
1.00
0.75
0.50
0.25
0.00
0.0
0.4
0.8
1.2
Fitness
Spearman's corr: 0.99
Emergence date
2023
2022
2021
2020

CoVFit-CLI

- The CoVFit models can be used as a stand-alone command line tool to predict the viral fitness of SARS-CoV-2 spike protein sequences in fasta format. Download link and instructions are available here.

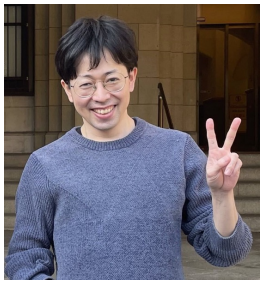
https://github.com/TheSatoLab/CoVFit

Integrative modeling of seasonal influenza evolution via AI-powered antigenic cartography

Jumpei Ito, Shusuke Kawakubo, Hiroaki Unno, Adam Strange, Spyros Lytras, Kaho Okumura, Alice Lilley, Ruth Harvey, Nicola Lewis, Kei Sato

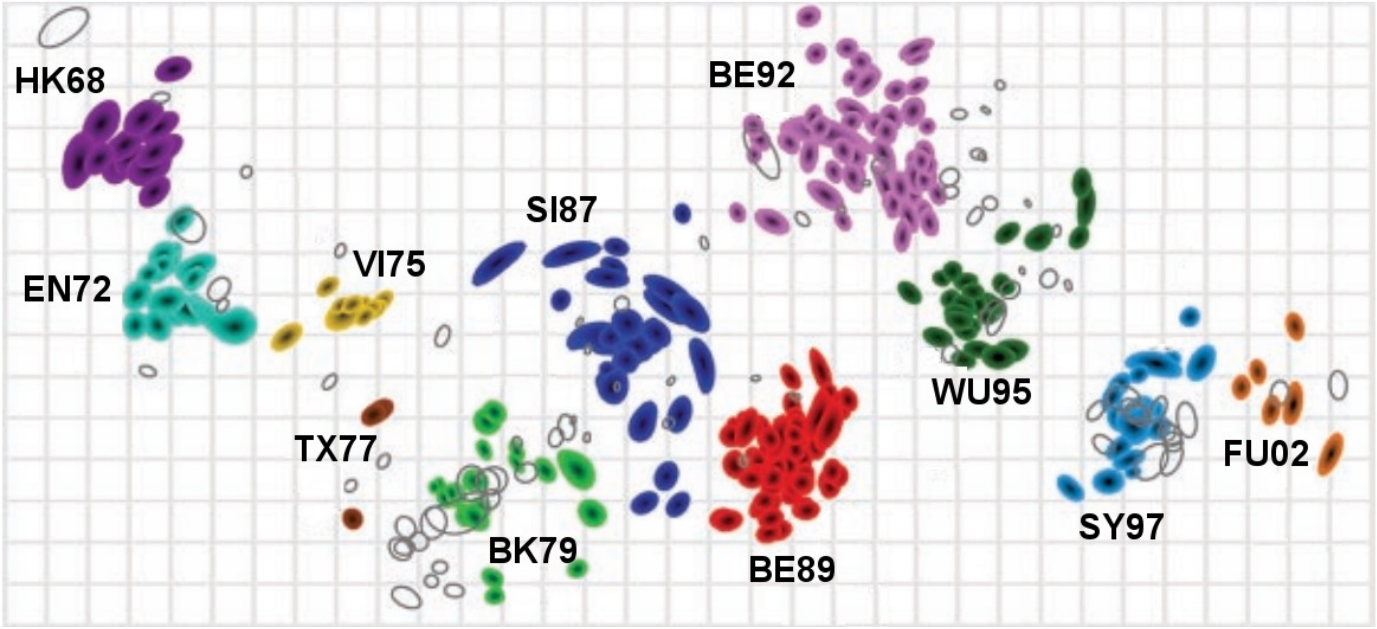
doi: https://doi.org/10.1101/2025.08.04.668423

Shusuke Kawakubo



Jumpei Ito

Predicting antigenic cartography coordinates of H3N2 human influenza



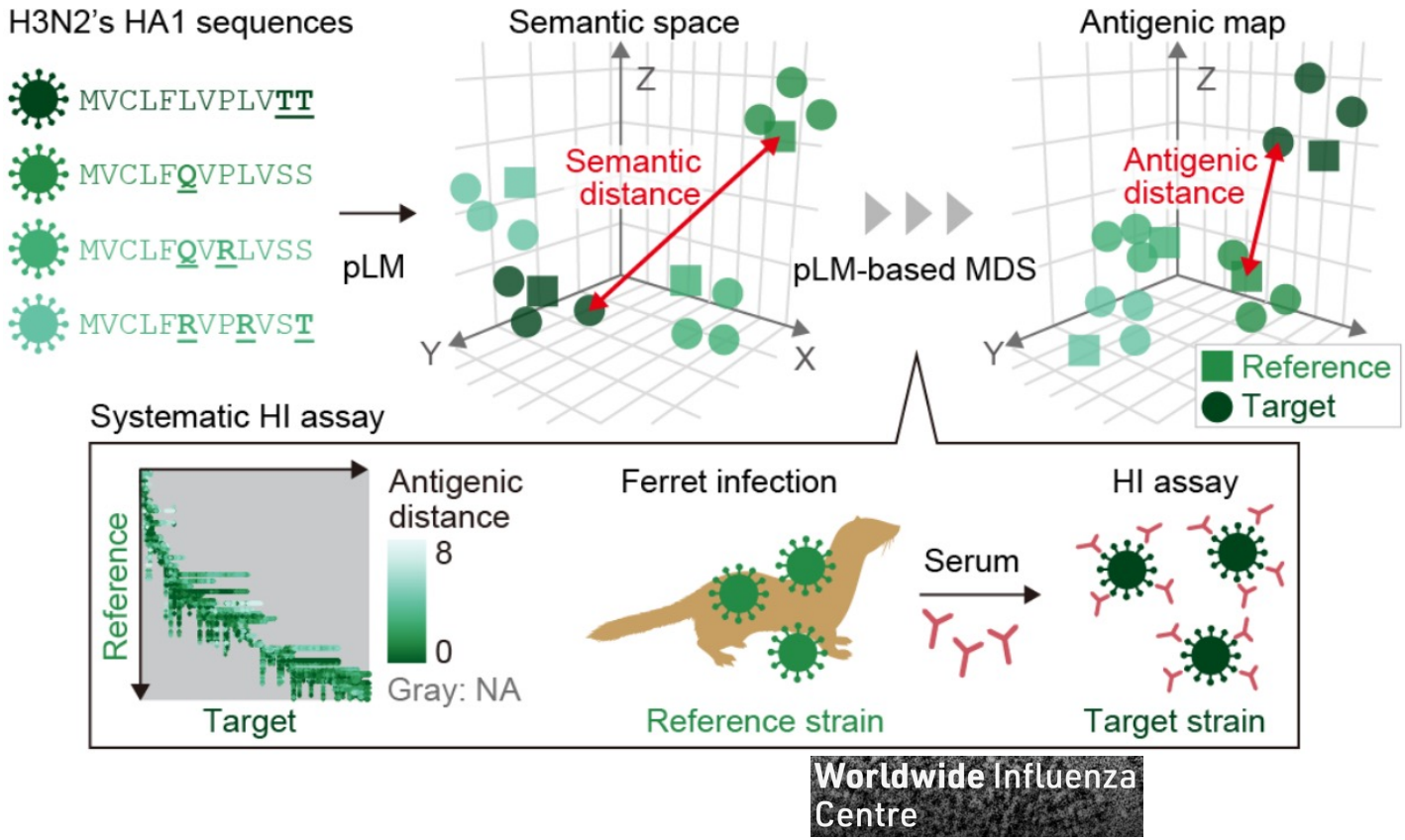
Smith *et al.* (2004) *Science* 305:5682

PLANT
(Protein Language Model for Antigenic cartography)

Integrative modeling of seasonal influenza evolution via AI-powered antigenic cartography

Jumpei Ito, Shusuke Kawakubo, Hiroaki Unno, Adam Strange, Spyros Lytras, Kaho Okumura, Alice Lilley, Ruth Harvey, Nicola Lewis, Kei Sato

doi: https://doi.org/10.1101/2025.08.04.668423



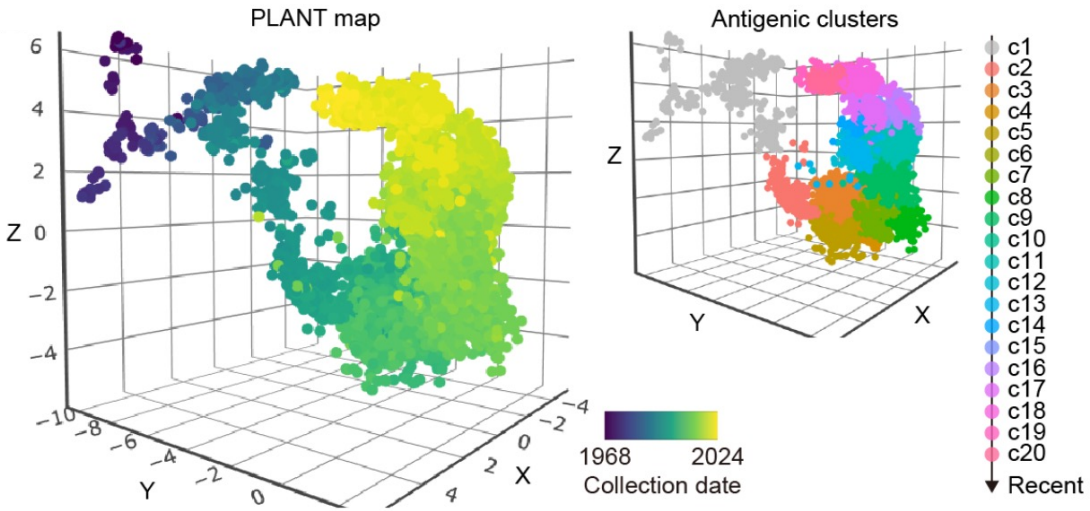
Worldwide Influenza Centre

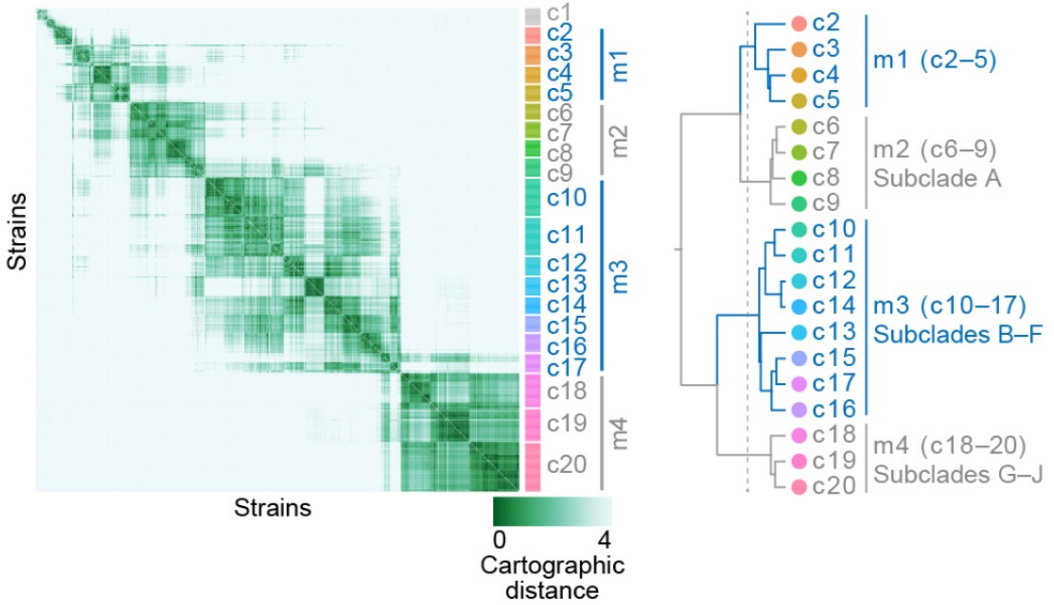
PLANT
(Protein Language Model for Antigenic cartography)

Integrative modeling of seasonal influenza evolution via AI-powered antigenic cartography

Jumpei Ito, Shusuke Kawakubo, Hiroaki Unno, Adam Strange, Spyros Lytras, Kaho Okumura, Alice Lilley, Ruth Harvey, Nicola Lewis, Kei Sato

doi: https://doi.org/10.1101/2025.08.04.668423





Trained model

The PLANT model, trained on data up to the 2024 Southern Hemisphere season (full model), is available on the Hugging Face repository: TheSatoLab-UTokyo/PLANT

Altough the original model used in the preprint, please use the fixed model with improved performance in **variants/PLANT_fixed**.

Core component of PLANT

Please refer to **src/plant/model.py** if you are interested in the implementation of PLANT and its pLM-DMS.

Google Colab notebook

A Colab notebook for embedding your sequences of interest onto the antigenic map constructed by the full model is available at the following link:

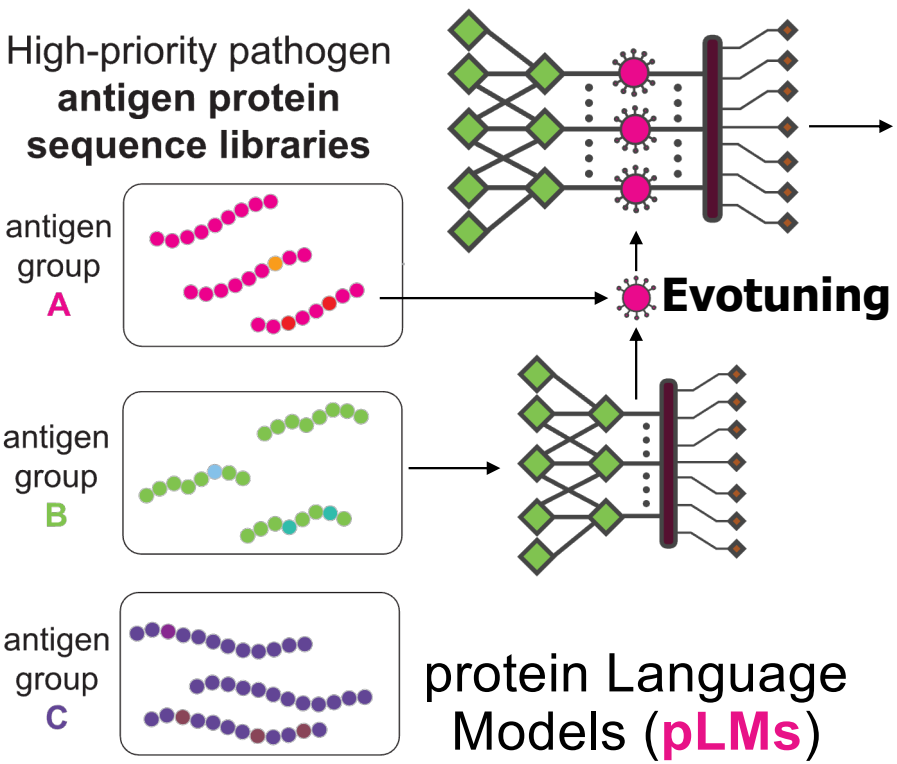
Open in Colab

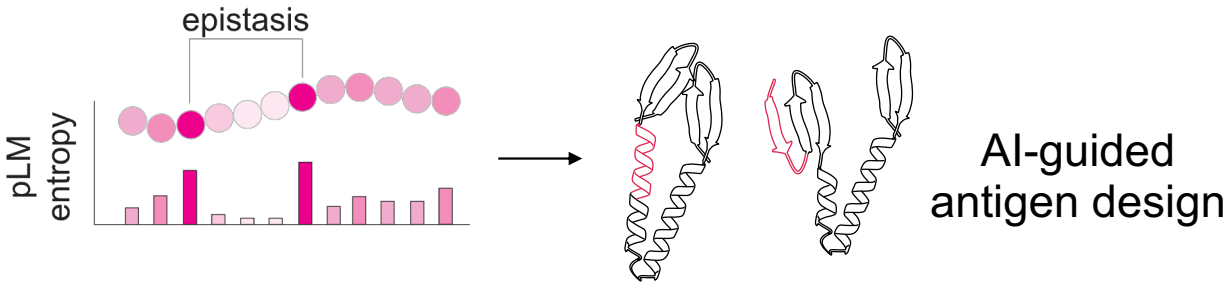
https://github.com/TheSatoLab/PLANT

https://colab.research.google.com/drive/1sLE3ysElImtxBBIzlGHlFTdDo8_O5aoY

AI for vaccine development

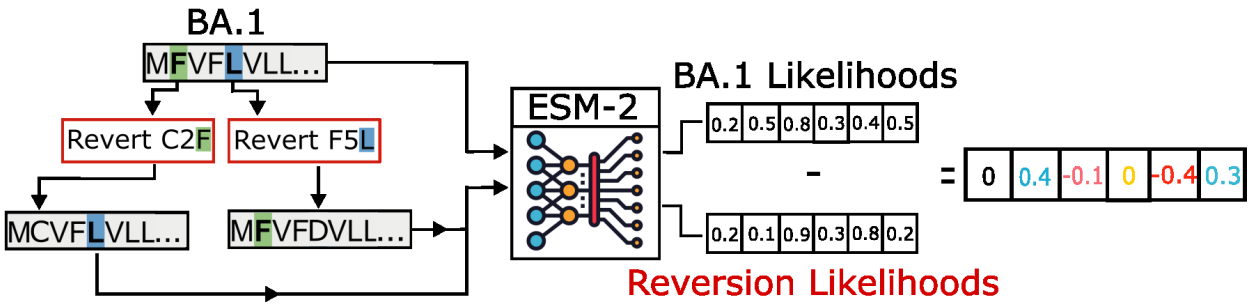
Can we use these AI-based technologies to improve how we make vaccines, by focusing on different antigens?





Antigen Evolution Design Laboratory

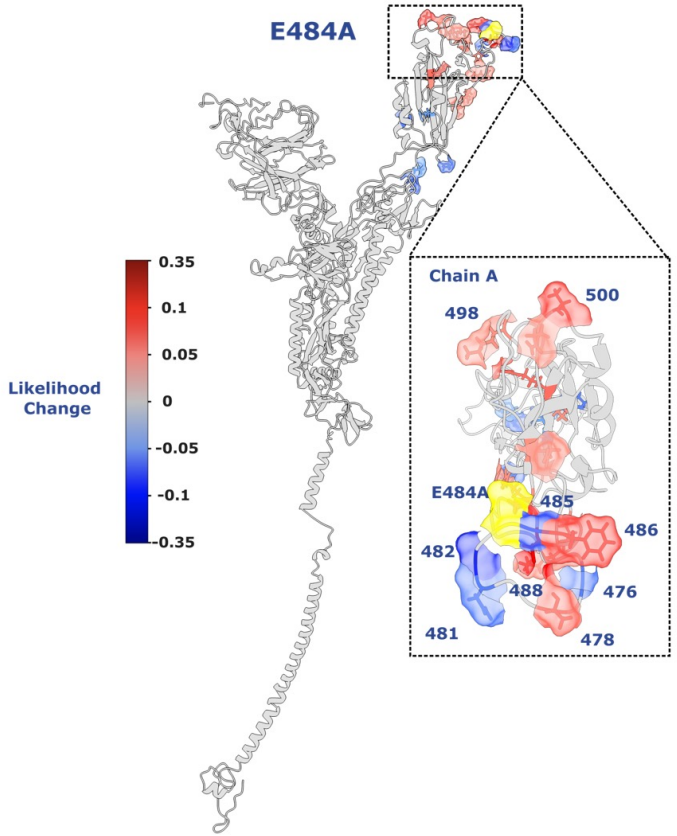
INSTITUT Pasteur

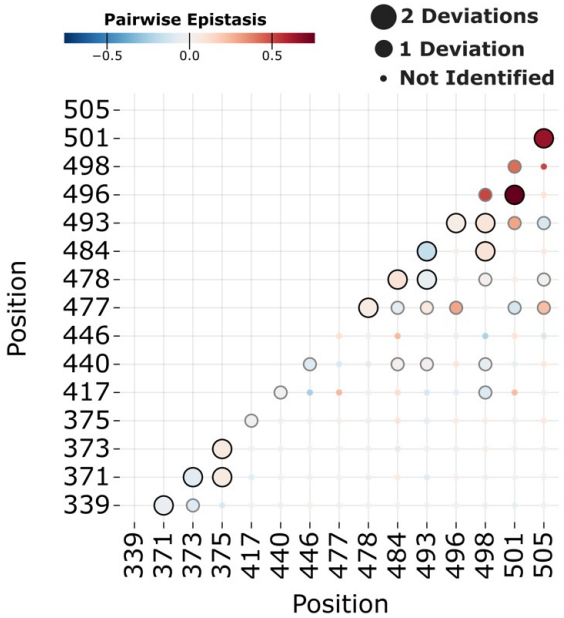


From single-sequences to evolutionary trajectories: protein language models capture the evolutionary potential of SARS-CoV-2

Kieran D. Lamb, Joseph Hughes, Spyros Lytras, Francesca Young, Orges Koci, James C. Herzig, Simon C. Lovell, Joe Grove, Ke Yuan ✉ & David L. Robertson ✉

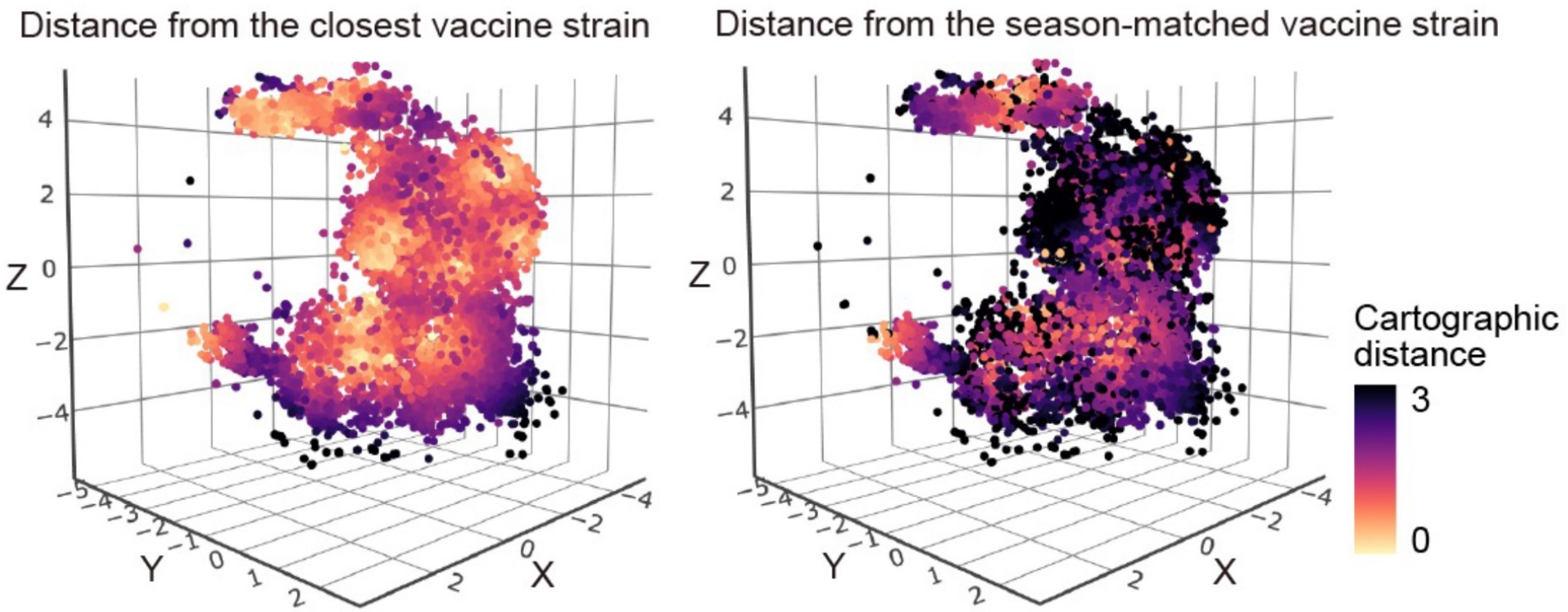
Nature Communications , Article number: (2026) | Cite this article





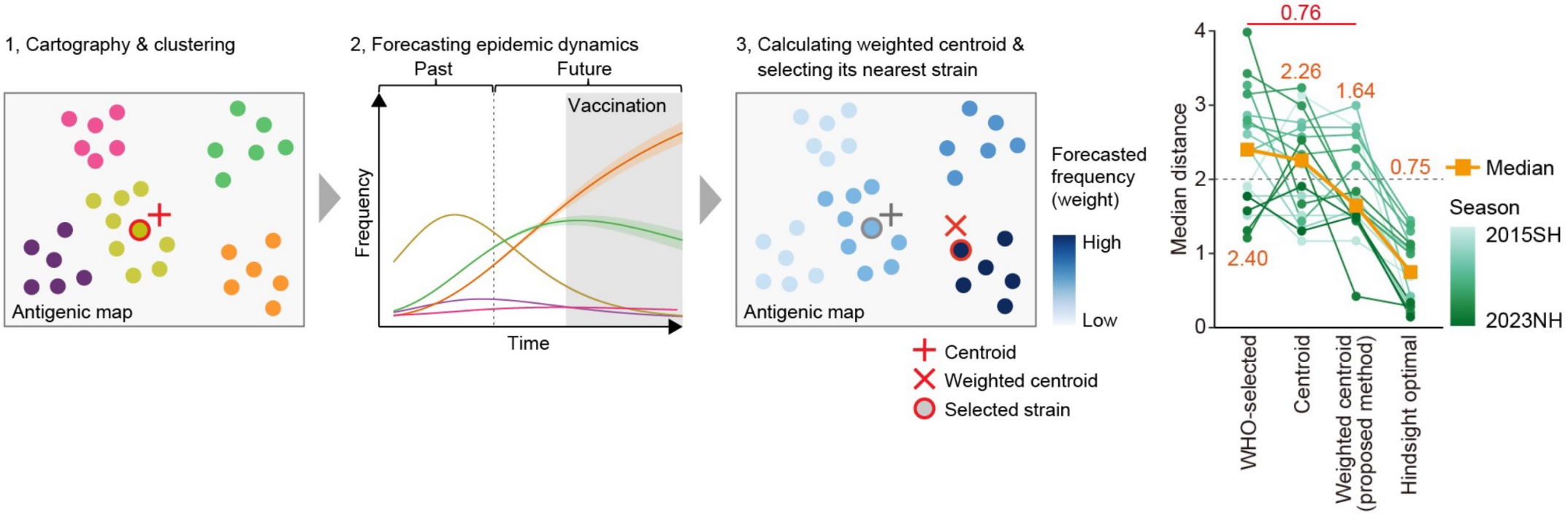
AI for vaccine development

Improving the ways that we currently use for selecting vaccine strains



AI for vaccine development

Improving the ways that we currently use for selecting vaccine strains



Assessing flu HA site variability with pLM entropy

Inferring context-specific site variation with evotuned protein language models

Spyros Lytras ✉, Adam Strange, Jumpei Ito ✉, Kei Sato ✉

NAR Genomics and Bioinformatics, Volume 8, Issue 1, March 2026, lqag018,
https://doi.org/10.1093/nargab/lqag018
Published: 09 February 2026 **Article history** ▾

https://github.com/spyros-lytras/plm_entropy

Predicting virus fitness with CoVFit

A protein language model for exploring viral fitness landscapes

Jumpei Ito ✉, Adam Strange, Wei Liu, Gustav Joas, Spyros Lytras, The Genotype to Phenotype Japan (G2P-Japan) Consortium & Kei Sato ✉

Nature Communications **16**, Article number: 4236 (2025) | Cite this article

https://github.com/TheSatoLab/CoVFit

Predicting virus antigenicity with PLANT

Integrative modeling of seasonal influenza evolution via AI-powered antigenic cartography

Jumpei Ito, Shusuke Kawakubo, Hiroaki Unno, Adam Strange, Spyros Lytras, Kaho Okumura, Alice Lilley, Ruth Harvey, Nicola Lewis, Kei Sato

doi: https://doi.org/10.1101/2025.08.04.668423

https://github.com/TheSatoLab/PLANT

Questions?

From June 2026…

Antigen Evolution Design Laboratory

INSTITUT Pasteur

Shusuke Kawakubo

Jumpei Ito

VIRUS INFORMATICS ITO Laboratory

Adam Strange

AI in Pandemic Preparedness and Vaccine development

Antigen Surveillance: from Evolution to Immune Escape