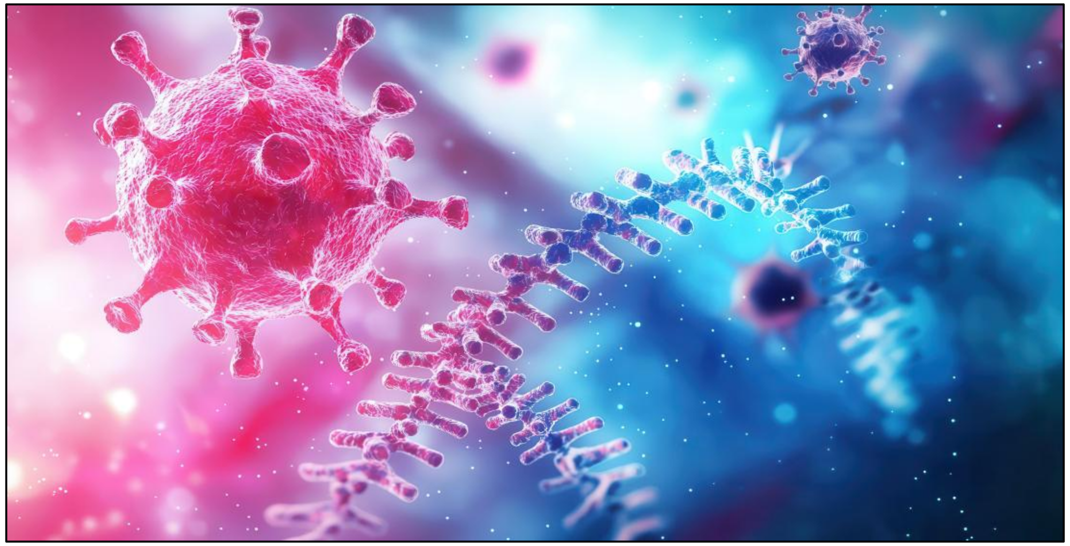
Learning the language of viral evolution and escape[1]

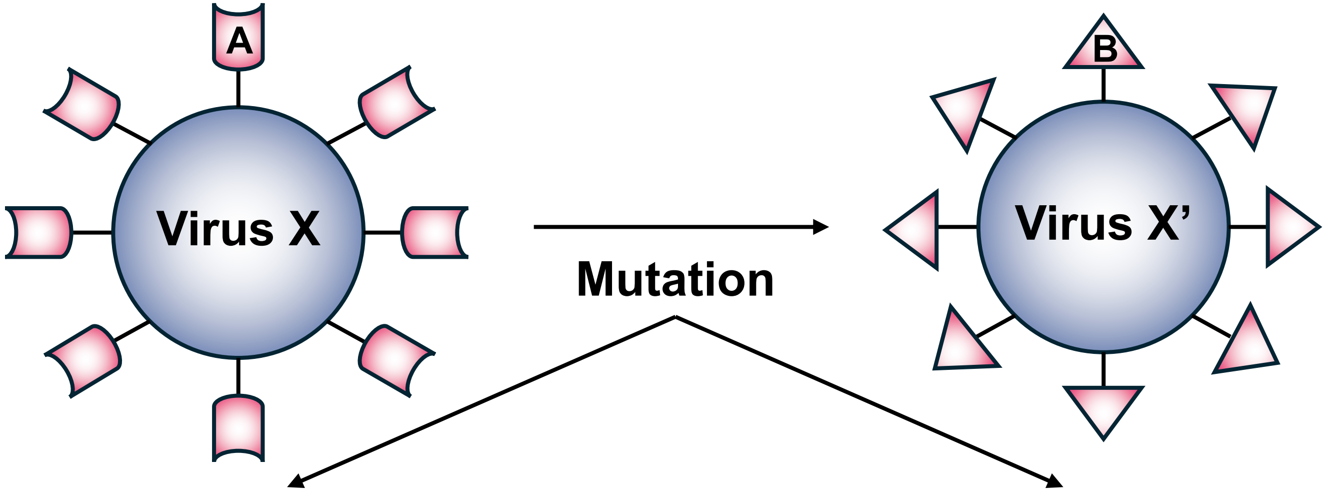


Hie, B., Zhong, E. D., Berger, B. & Bryson, B. *Science* 371, 284–288 (2021).

Conor Warren

COS 598L 3/17/2026

Central Idea



Unrecognizable to host immune system!

Significantly Transform Protein ↕ Transform Meaning of Sentence

Maintain Protein's Fitness ↕ Maintain Grammar of Sentence

Natural Language Analogs

The cat ran quickly → The cat ran slowly

Outline

1. Background

1. Viruses & Viral Escape
2. Experimental Escape Modeling
3. Computational Escape Modeling
4. Analogy to Language Modeling

2. Methodology

1. Overview of Approach
2. Viral Protein Language Model
3. Viral Escape Prediction
4. Experiments & Implementation

3. Results

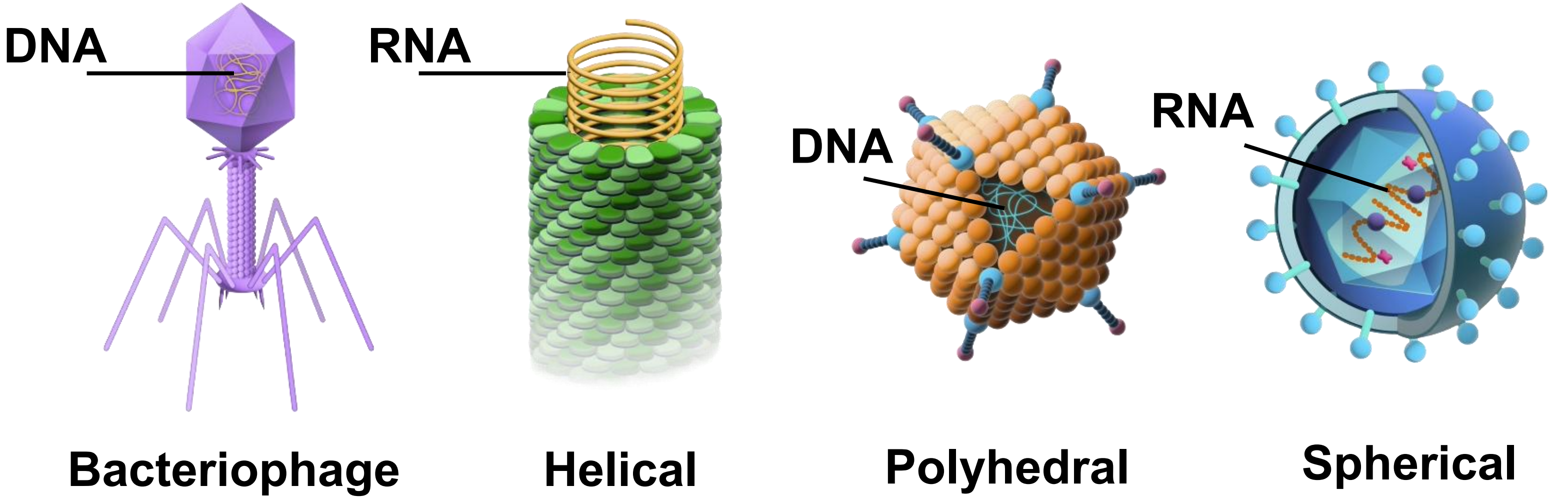
1. Semantic Embedding Clustering
2. Semantics & Grammaticality
3. Escape Prediction Performance
4. Structural Localization of Escape

4. Discussion

1. Key Findings & Takeaways
2. Strengths & Limitations
3. Future Directions
4. Concluding Remarks

Viruses & Viral Escape

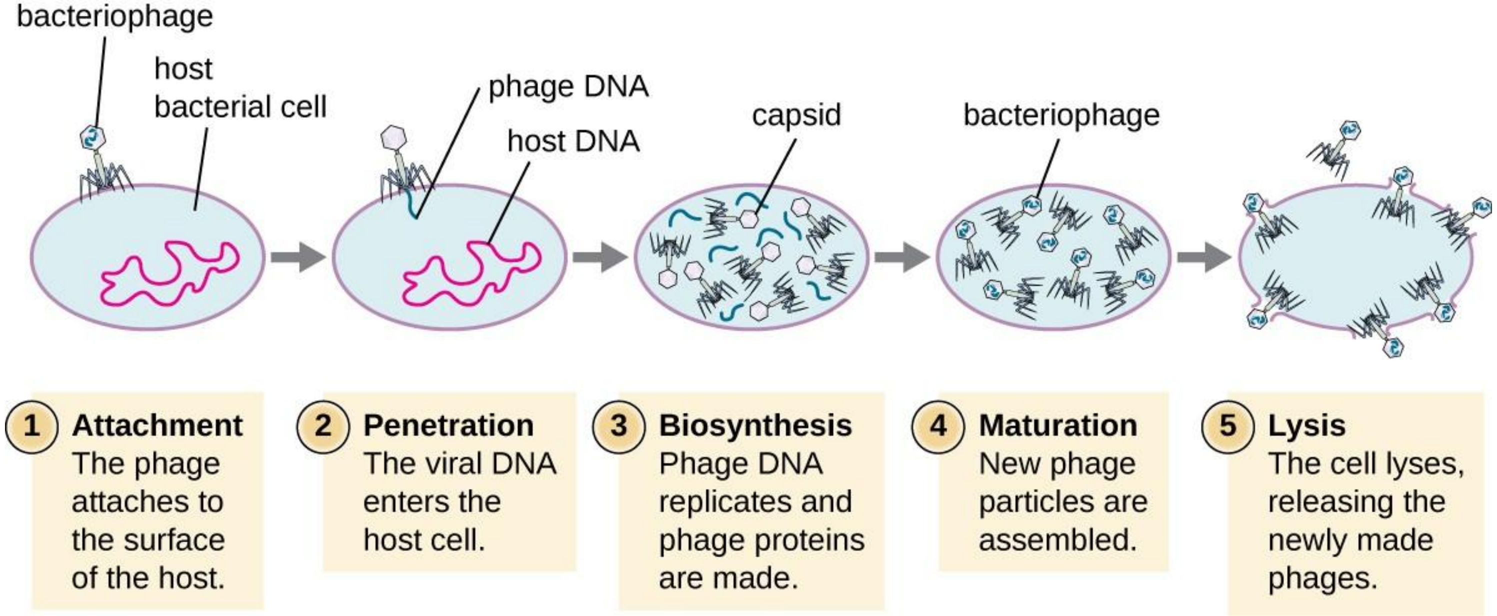
Viruses are biological agents that consist of genetic material and a protective protein shell[2,3]



Adapted from [4,5]

Viruses & Viral Escape

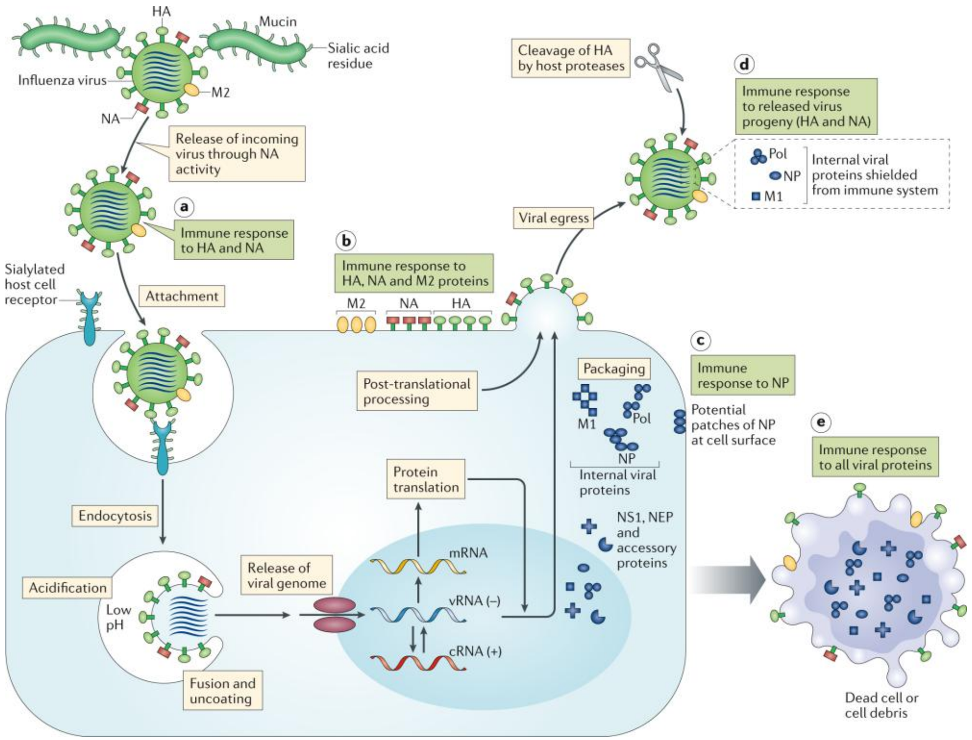
They are characterized by their hijacking and exploitation of host cells for reproduction and protein synthesis[6,7]


bacteriophage
host bacterial cell
phage DNA
host DNA
capsid
bacteriophage
1 Attachment
The phage attaches to the surface of the host.
2 Penetration
The viral DNA enters the host cell.
3 Biosynthesis
Phage DNA replicates and phage proteins are made.
4 Maturation
New phage particles are assembled.
5 Lysis
The cell lyses, releasing the newly made phages.

From [8]

Viruses & Viral Escape

The “success” of a virus is tied to its capacity to evade detection by host immune systems[9, 10]


HA
Mucin
Sialic acid residue
Influenza virus
M2
NA
Release of incoming virus through NA activity
a
Immune response to HA and NA
Sialylated host cell receptor
Attachment
b
Immune response to HA, NA and M2 proteins
M2
NA
HA
Cleavage of HA by host proteases
d
Immune response to released virus progeny (HA and NA)
Pol
NP
M1
Internal viral proteins shielded from immune system
Viral egress
Post-translational processing
Packaging
M1
Pol
NP
Internal viral proteins
c
Immune response to NP
Potential patches of NP at cell surface
e
Immune response to all viral proteins
Protein translation
Endocytosis
NS1, NEP and accessory proteins
mRNA
Release of viral genome
Acidification
Low pH
vRNA (–)
cRNA (+)
Fusion and uncoating
Dead cell or cell debris

From [11]

Viruses & Viral Escape

Viruses have thus evolved intricate strategies for this purpose[9]

DNA-based Viruses

Camouflage

Wrap payload in host membrane

Sabotage

Interfere with host immune system

RNA-based Viruses

Speed

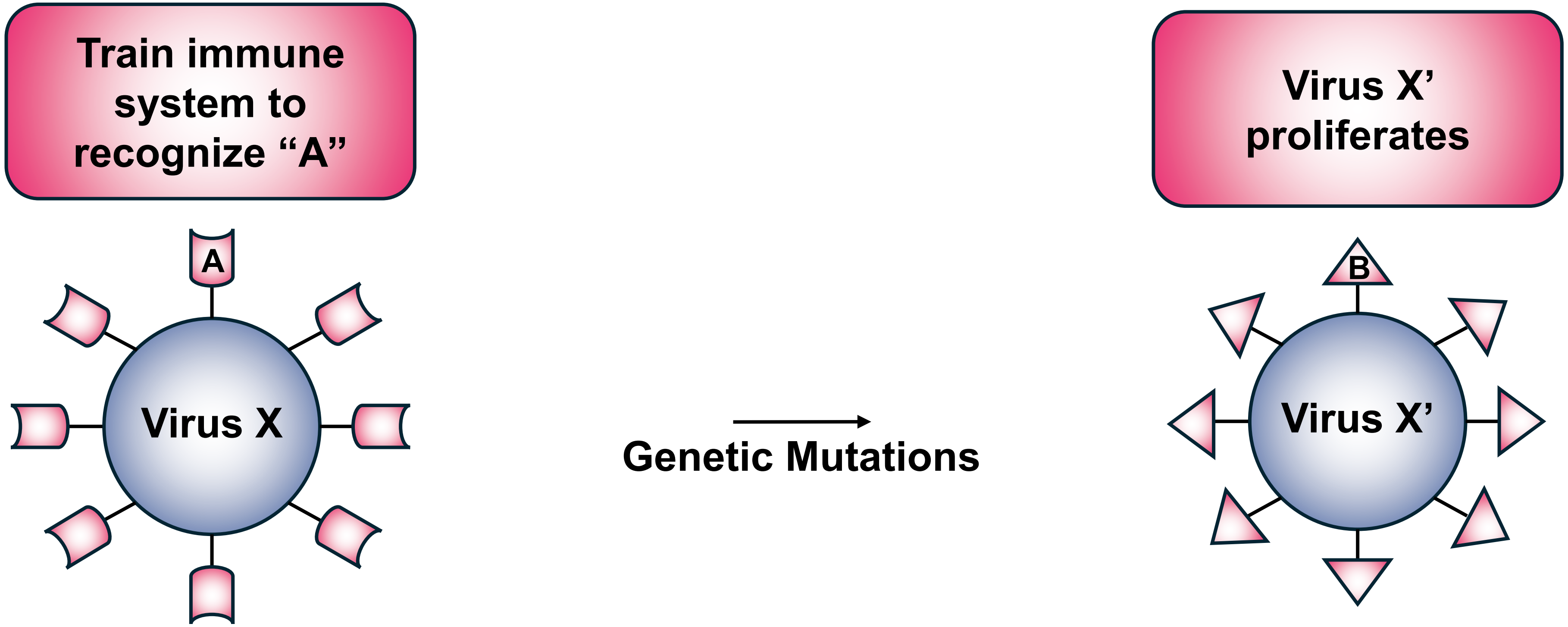
Overwhelm host immune system through rapid replication

Transformation/Shape Change

Accumulate mutations that prevent detection

Viruses & Viral Escape

The capacity of a virus to elude immune defenses has thwarted the development of effective vaccines and treatments for viral infections



Train immune system to recognize “A”

Genetic Mutations

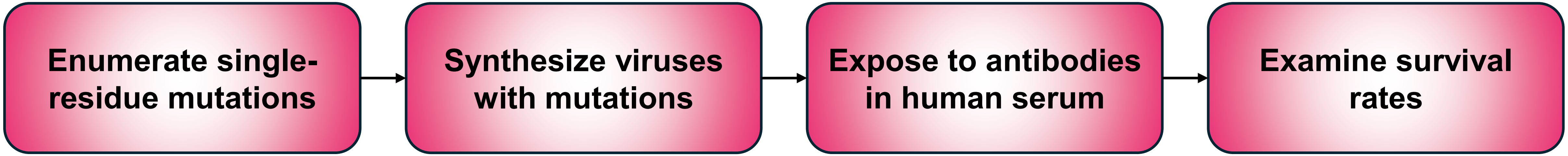
Virus X’ proliferates

Viruses & Viral Escape

Against this backdrop, understanding which mutations are likely to promote viral escape is critically important

Experimental Escape Modeling

The standard approach to viral escape modeling involves enumerating single-residue mutations and measuring virus survival *in vitro*[12,13]



Enumerate single-residue mutations

Synthesize viruses with mutations

Expose to antibodies in human serum

Examine survival rates

Mutants that survive are likely to exhibit viral escape in vivo

This setup produces informative results,
but it is arduous and does not scale well

Computational Escape Modeling

Accordingly, computational methods have been developed to model viral escape through protein evolution *in silico*

Fitness

Determine natural likelihood of amino acid sequences

Likelihood ≈ Biological Fitness

OR

Functional/Semantic Similarity

Generate vector representations of proteins

Measure similarity between mutant and wild-type representations

Might it be helpful to consider both fitness and functional/semantic similarity?

Analogy to Language Modeling

To motivate their to approach this task, Hie et al. draw an analogy between natural language modeling and viral escape modeling

Natural Language Modeling

Sentences are sequences of words

Sentence alterations may significantly change the meaning of a sentence

But these alterations must be grammatical to preserve linguistic validity

Viral Escape Modeling

Proteins are sequences of amino acids

Viral mutations must significantly change a virus to promote escape

But these mutations must obey biological rules to preserve fitness

If this analogy holds, then language models may be effective at modeling viral escape!

Outline

1. Background

1. Viruses & Viral Escape
2. Experimental Escape Modeling
3. Computational Escape Modeling
4. Analogy to Language Modeling

2. Methodology

1. Overview of Approach
2. Viral Protein Language Model
3. Viral Escape Prediction
4. Experiments & Implementation

3. Results

1. Semantic Embedding Clustering
2. Semantics & Grammaticality
3. Escape Prediction Performance
4. Structural Localization of Escape

4. Discussion

1. Key Findings & Takeaways
2. Strengths & Limitations
3. Future Directions
4. Concluding Remarks

Overview of Approach

The central hypothesis of this project is as follows:

The viral mutations likely to cause viral escape are those that:

1. **“Preserve viral infectivity”**
2. **“Cause a virus to look different to the immune system”**

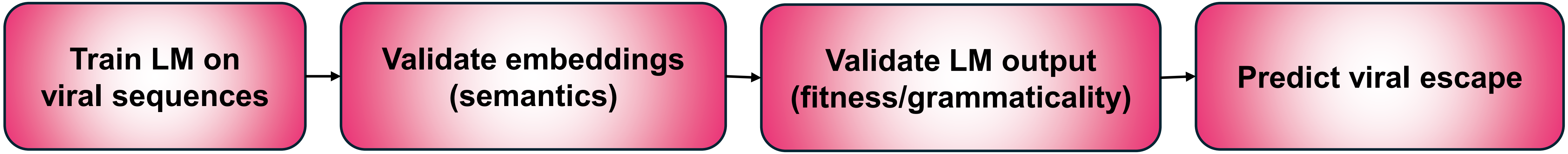
This is analogous to sentence alterations that:

1. **“Preserve a sentence’s grammaticality”**
2. **“Change [a sentence’s] meaning”**

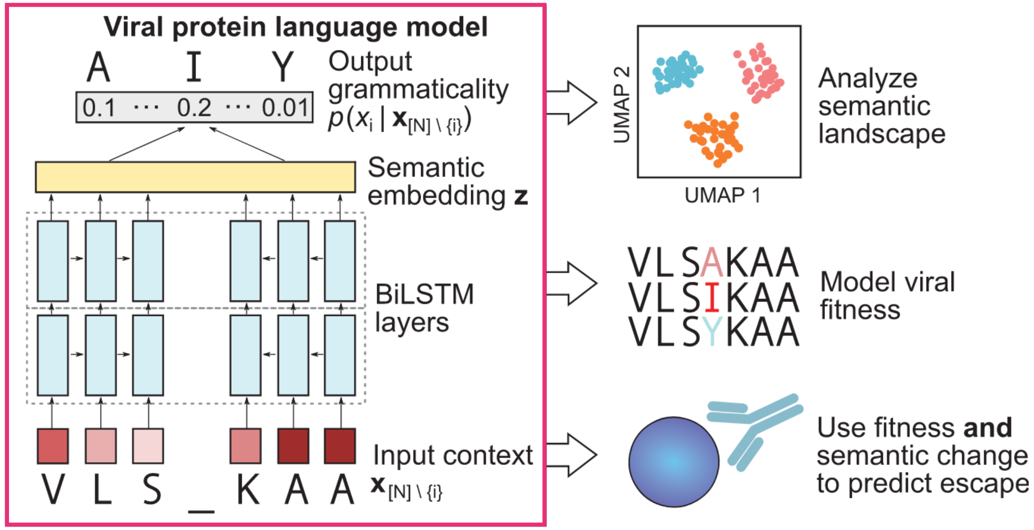
The viral analogs of grammaticality and semantic meaning can be estimated with a language model trained on viral sequence data

Overview of Approach

To test this hypothesis, Hie et al. adopt the following approach:



Train LM on viral sequences → **Validate embeddings (semantics)** → **Validate LM output (fitness/grammaticality)** → **Predict viral escape**


Viral protein language model
A I Y
Output grammaticality
0.1 ··· 0.2 ··· 0.01
$p(x_i \mid \mathbf{x}_{[N]\setminus\{i\}})$
Semantic embedding **z**
BiLSTM layers
Input context
V L S _ K A A
$\mathbf{x}_{[N]\setminus\{i\}}$
UMAP 2
UMAP 1
Analyze semantic landscape
VLSAKAA
VLSIKAA
VLSYKAA
Model viral fitness
Use fitness **and** semantic change to predict escape

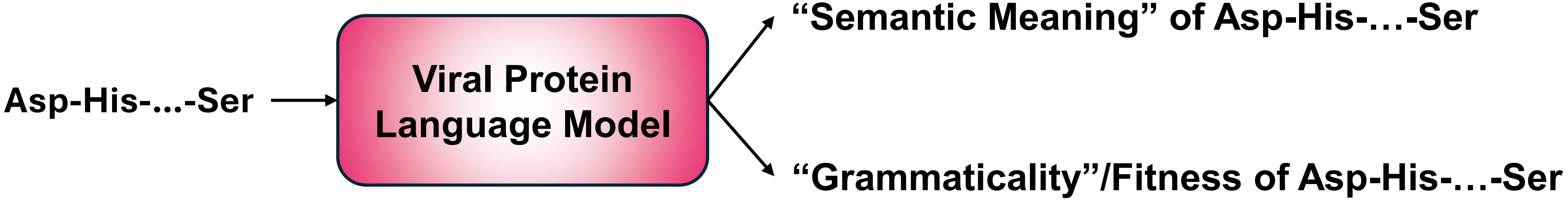
Viral Protein Language Model

To understand the viral protein language model, let's first formulate the problem it solves here

Given the amino acid sequence of a viral protein $\boldsymbol{x} = [\boldsymbol{x_1}, \dots, \boldsymbol{x_n}]$:

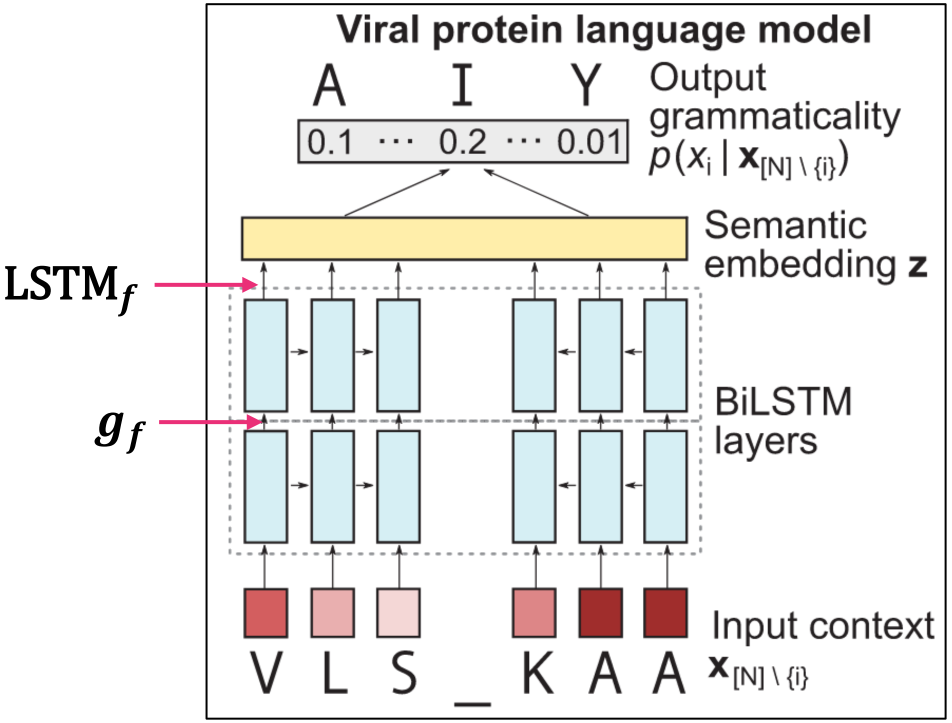
1. **Find a semantic embedding $\boldsymbol{z} = \boldsymbol{f_s}(\boldsymbol{x})$, where $\boldsymbol{f_s}(\boldsymbol{x})$ maps "semantically similar" proteins to similar $\boldsymbol{K}$-dimensional coordinates**

2. **Determine the "grammaticality" of $\boldsymbol{x}$**


Asp-His-...-Ser
Viral Protein Language Model
"Semantic Meaning" of Asp-His-…-Ser
"Grammaticality"/Fitness of Asp-His-…-Ser

Viral Protein Language Model

Stripping away the abstraction, a bidirectional LSTM is employed to achieve these functionalities



<u>Objective: Masked Amino Acid Prediction</u>

Predict each masked amino acid based on prior and future context

<u>Semantic Embedding</u>

$$z_i = [\mathrm{LSTM}_f(g_f(x_1, \dots, x_{i-1})^T, \dots, \mathrm{LSTM}_r(g_r(x_{i+1}, \dots, x_N)^T]$$

<u>Grammaticality: Likelihood</u>

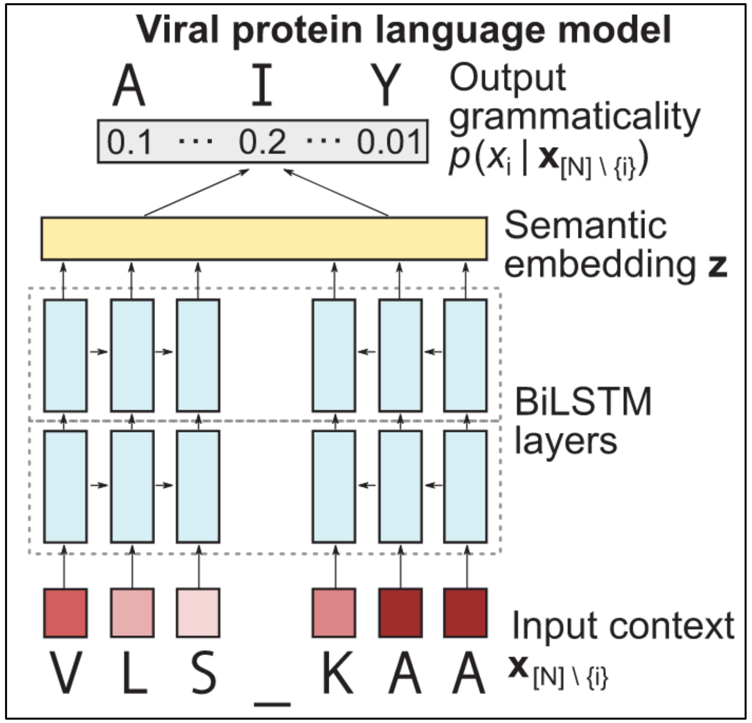
$$P(x_i | z_i)$$

Probability distribution over all possible amino acids for position i given prior and future context

Correlated with grammaticality in natural language

Viral Escape Prediction

How the semantic embedding and grammaticality of the viral protein sequence are utilized for escape prediction requires elaboration


Viral protein language model
A
I
Y
Output grammaticality $p(x_i \mid \mathbf{x}_{[N] \setminus \{i\}})$
0.1 ··· 0.2 ··· 0.01
Semantic embedding $\mathbf{z}$
BiLSTM layers
Input context $\mathbf{x}_{[N] \setminus \{i\}}$
V L S _ K A A

Suppose a mutation $\tilde{x}_i$ occurs at position i, resulting in the amino acid sequence $x[\tilde{x}_i] = [\ldots, \tilde{x}_i, \ldots]$

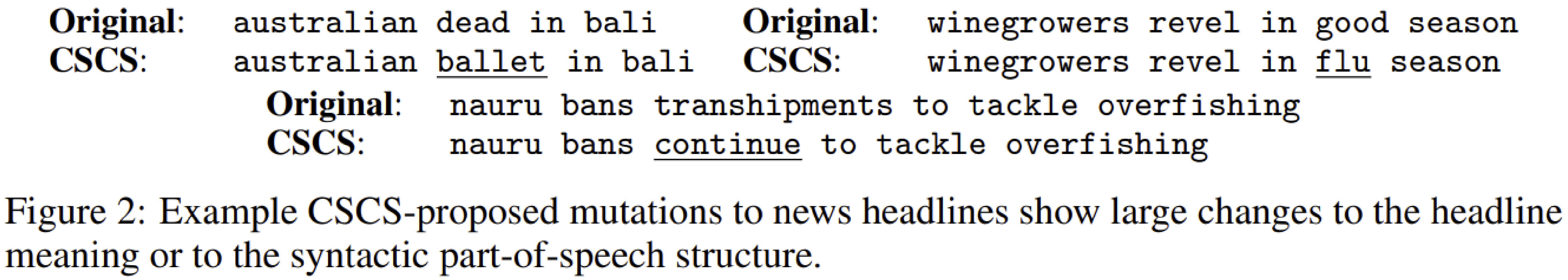
1. **Semantic Change**
   1. **Compute $z = f_s(x)$ and $\tilde{z} = f_s(x[\tilde{x}_i])$**
   2. **Compute $\Delta z[\tilde{x}_i] = ||z - \tilde{z}||$**

2. **Grammaticality**
   1. **Compute $\mathrm{P}(\mathbf{x_i} | \mathbf{x}_{[\mathbf{N}]\setminus\{\mathbf{i}\}})$**

3. **Constrained Semantic Change Search**
   1. **Compute $\alpha(\tilde{x}_i; x) = \Delta z[\tilde{x}_i] + \beta\, \mathrm{P}(\mathbf{x_i} | \mathbf{x}_{[\mathbf{N}]\setminus\{\mathbf{i}\}})$, where $\beta \in [0, \infty]$**

Viral Escape Prediction

Constrained Semantic Change Search Example: Natural Language



Original: `australian dead in bali`
CSCS: `australian ballet in bali`

Original: `winegrowers revel in good season`
CSCS: `winegrowers revel in flu season`

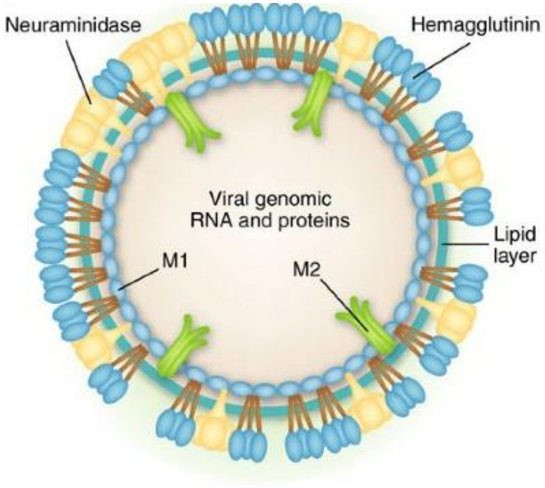
Original: `nauru bans transhipments to tackle overfishing`
CSCS: `nauru bans continue to tackle overfishing`

Figure 2: Example CSCS-proposed mutations to news headlines show large changes to the headline meaning or to the syntactic part-of-speech structure.

Experiments & Implementation

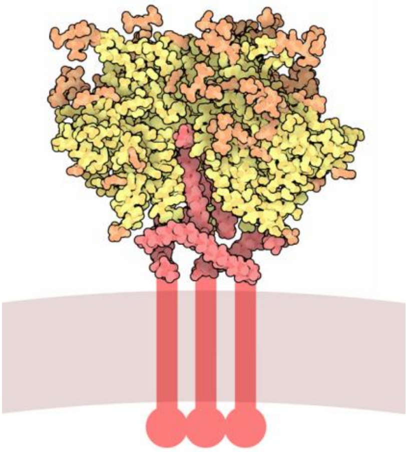
This pipeline is validated through experiments on three viral surface proteins that are pathologically important

Influenza A hemagglutinin (HA)



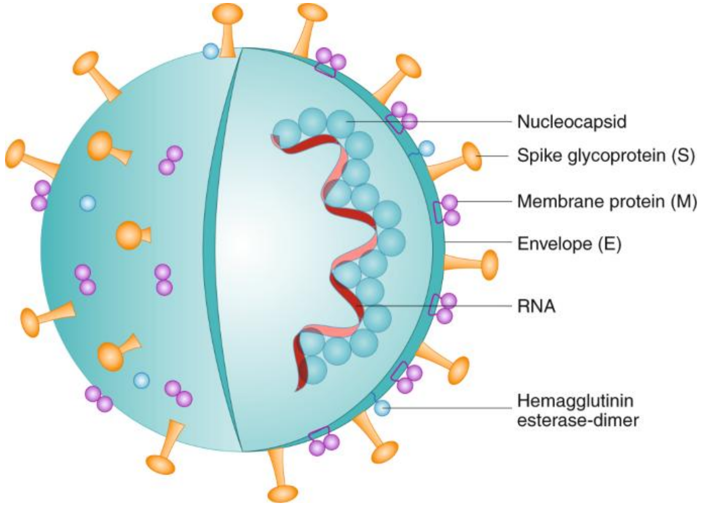
From [14]

HIV-1 Envelope Glycoprotein (Env)



From [15]

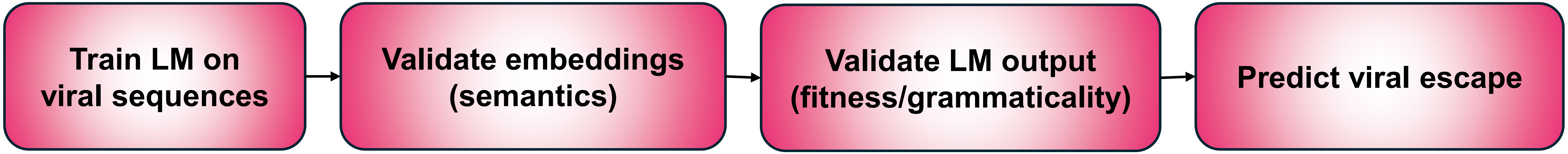
SARS CoV-2 Spike Glycoprotein (Spike)



From [16]

Experiments & Implementation

For each viral protein, the following steps are performed:



Train LM on viral sequences

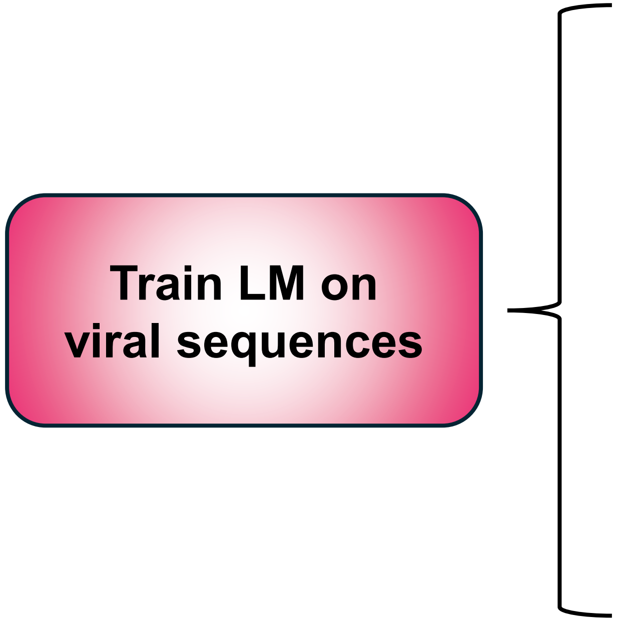
Validate embeddings (semantics)

Validate LM output (fitness/grammaticality)

Predict viral escape

Experiments & Implementation

A separate bidirectional LSTM is trained on real-world amino acid sequences corresponding to each viral protein of interest



Train LM on viral sequences

Train bidirectional LSTM for next amino acid prediction task on:

HA: 44,851 unique amino acid sequences

Env: 57,730 unique amino acid sequences

Spike: 4,172 unique amino acid sequences

Outline

1. Background

1. Viruses & Viral Escape
2. Experimental Escape Modeling
3. Computational Escape Modeling
4. Analogy to Language Modeling

2. Methodology

1. Overview of Approach
2. Viral Protein Language Model
3. Viral Escape Prediction
4. Experiments & Implementation

3. Results

1. Semantic Embedding Clustering
2. Semantics & Grammaticality
3. Escape Prediction Performance
4. Structural Localization of Escape

4. Discussion

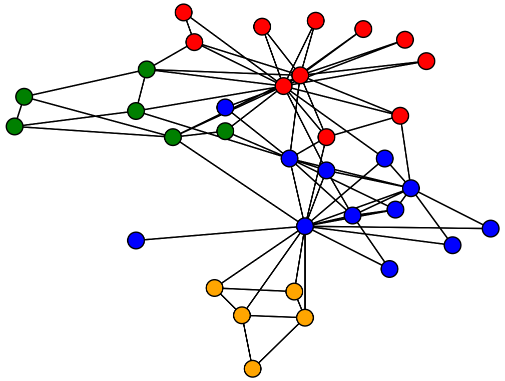
1. Key Findings & Takeaways
2. Strengths & Limitations
3. Future Directions
4. Concluding Remarks

Semantic Embedding Clustering

To ensure that viral protein language model captures semantic information, clustering analysis is performed

Validate embeddings (semantics)

1. **Construct KNN graph for sequence embeddings**
2. **Perform unsupervised clustering with Louvain community detection**
   1. **Evaluate cluster purity w.r.t sequence metadata labels**



From [17]

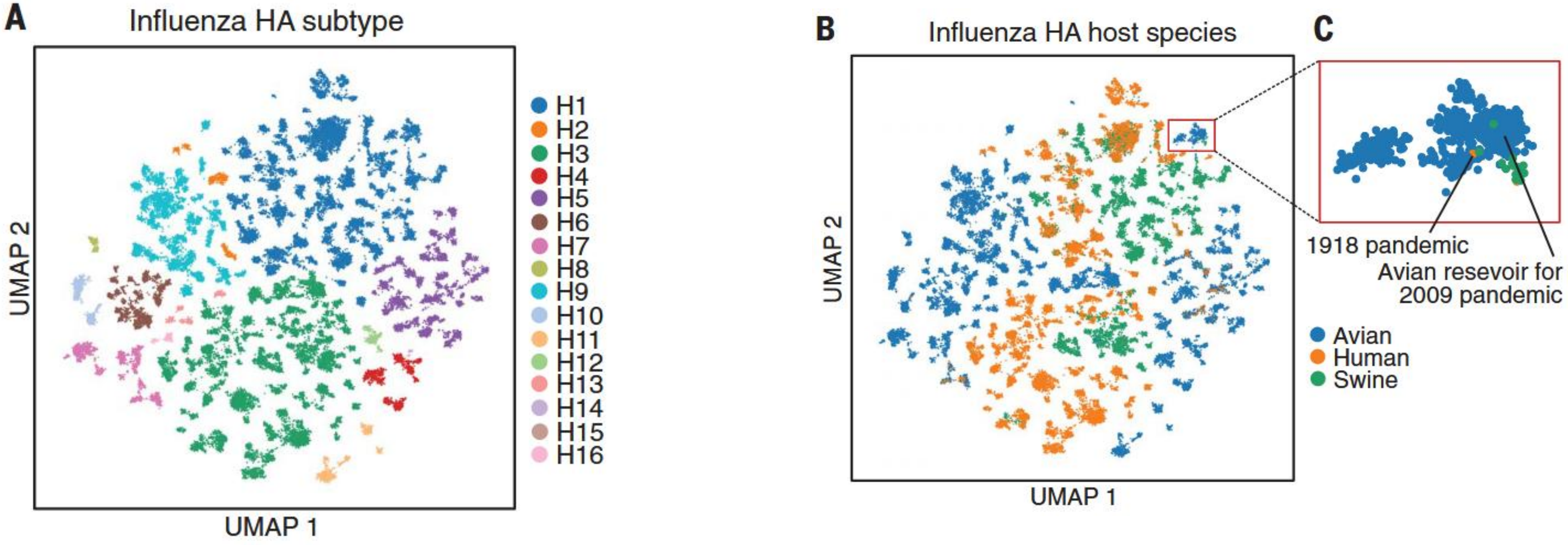
🔴 **80% Avian, …, 6% Feline**
10% Subtype A, …, 10% Subtype J

🔵 **10% Avian, …, 10% Feline**
90% Subtype A, …, 10% Subtype J

Helps identify semantic information encoded in clustered representations

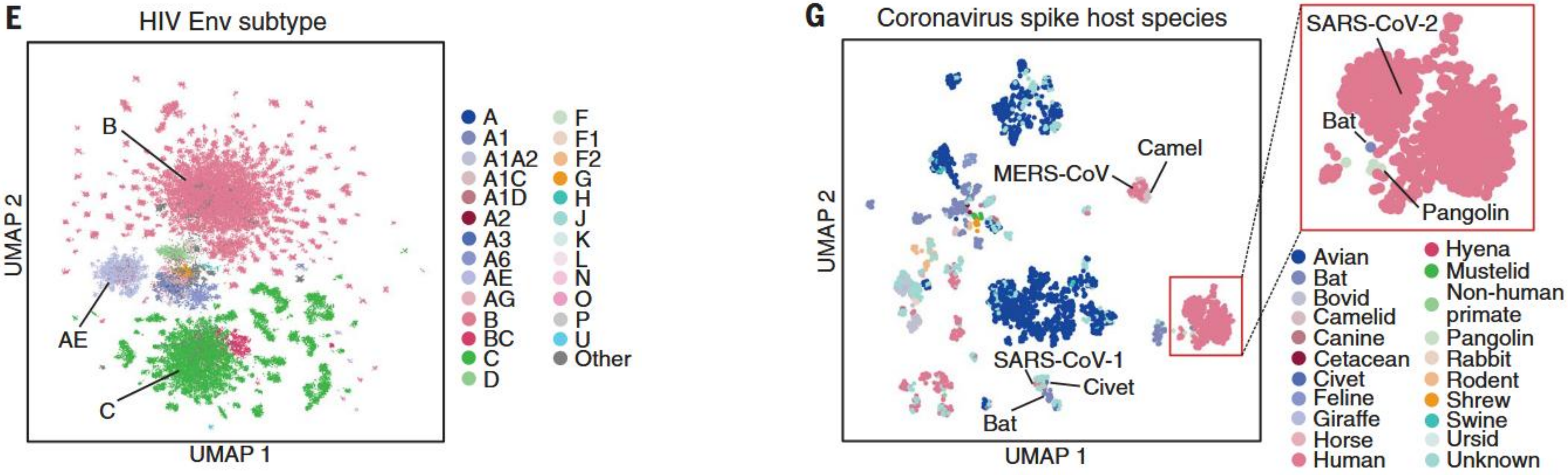
Semantic Embedding Clustering

The semantic embeddings of viral protein sequences encode subtype and host species information



Semantic Embedding Clustering

The semantic embeddings of viral protein sequences encode subtype and host species information



Experiments & Implementation

To assess viral fitness/grammaticality, the correlation is determined between the computed grammaticality and experimental metrics

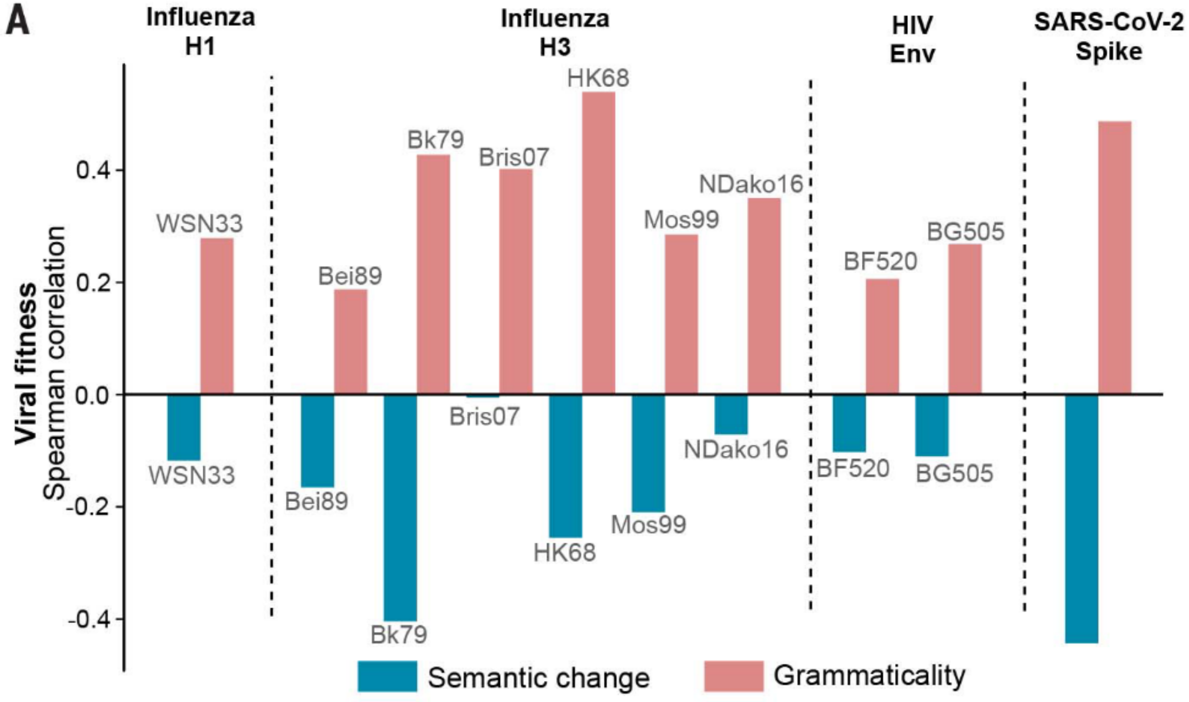
Validate LM output (fitness/grammaticality)

1. **Acquire datasets measuring:**
   1. **Replication fitness of single-residue mutations for HA and Env**
   2. **Binding affinity between Spike mutants and human ACE2**

2. **Compute Spearman correlation between computed grammaticality of mutant sequences and experimental measures**

Semantics & Grammaticality

Grammaticality and fitness are positively correlated; semantic change and fitness are negatively correlated



Experiments & Implementation

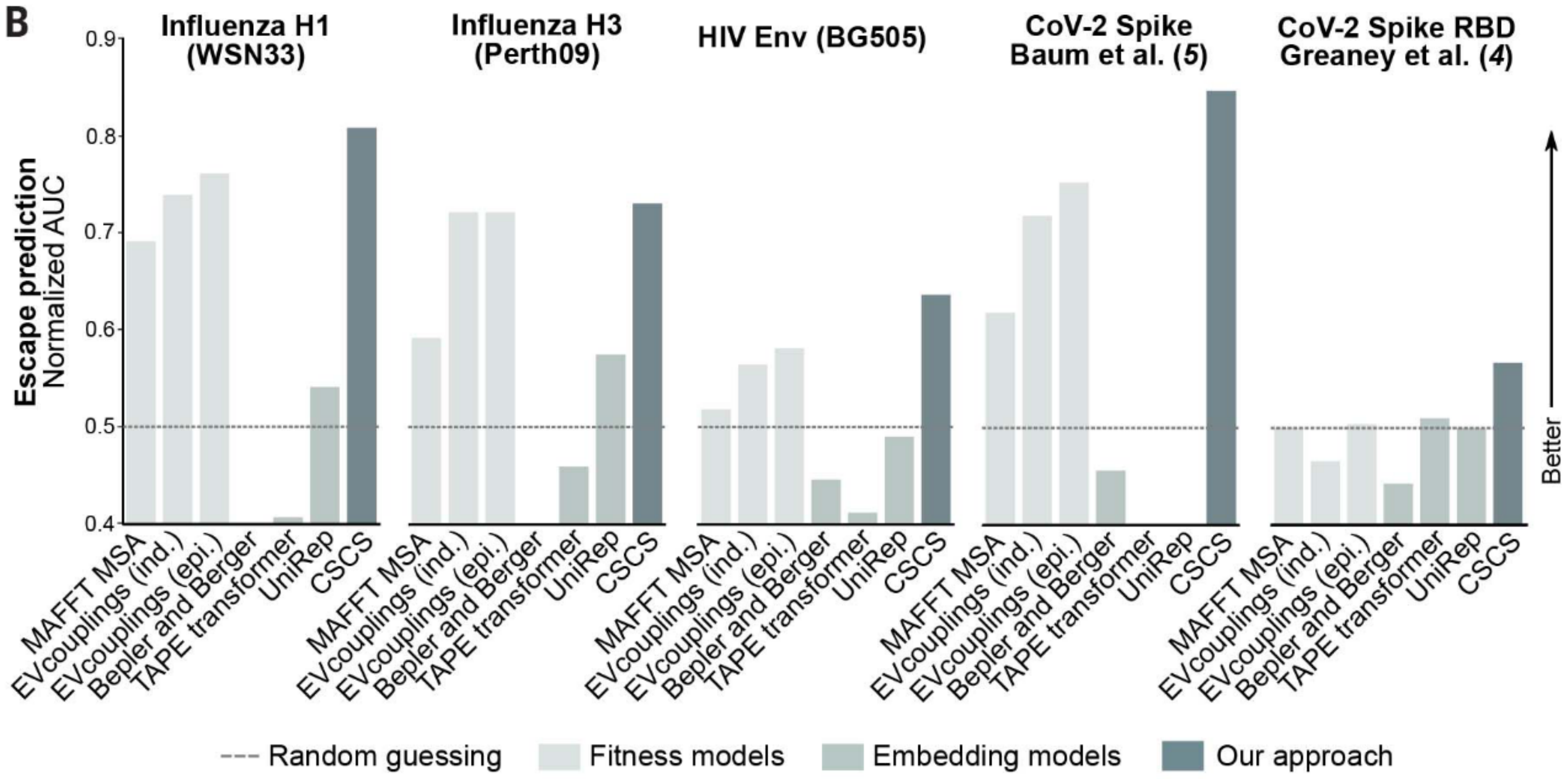
To assess viral fitness/grammaticality, the correlation is determined between the computed grammaticality and experimental metrics

Predict viral escape

1. **Acquire experimental escape modeling datasets that indicate which mutations promote viral escape**
2. **Perform CSCS to rank all possible mutations**
3. **Plot the top *n* CSCS mutants on x-axis and the number of these *n* mutants that were causal escape mutants; measure AUC**
    i. **What fraction of the top *n* CSCS mutants are causal escape mutants?**

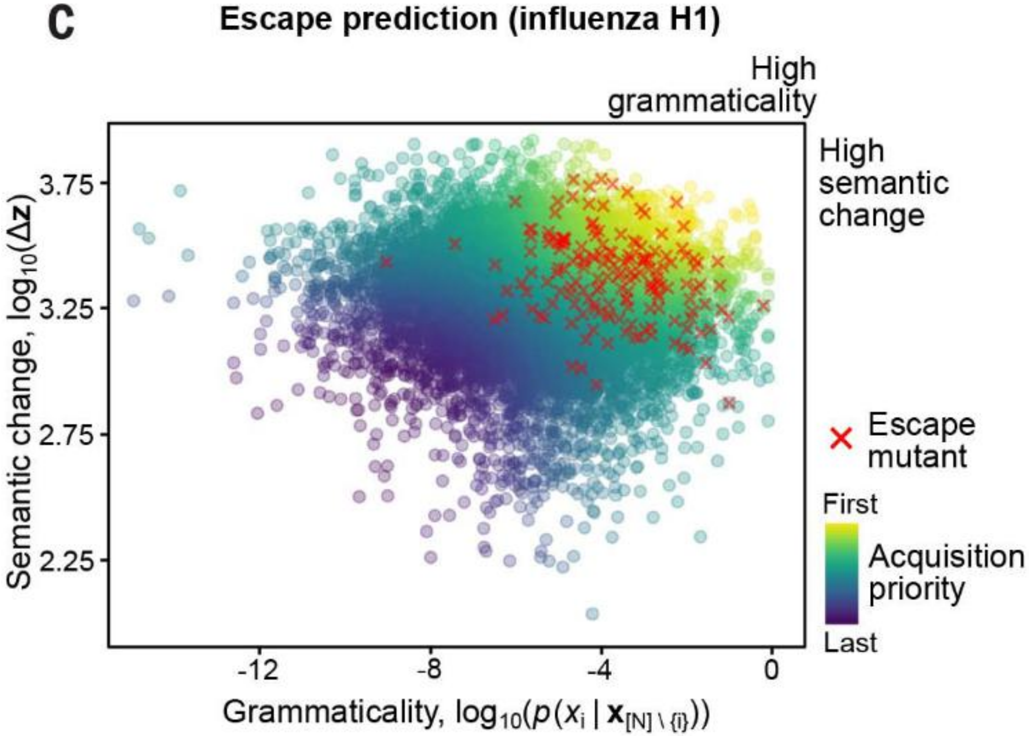
Escape Prediction Performance

CSCS outperforms other methods for viral escape prediction



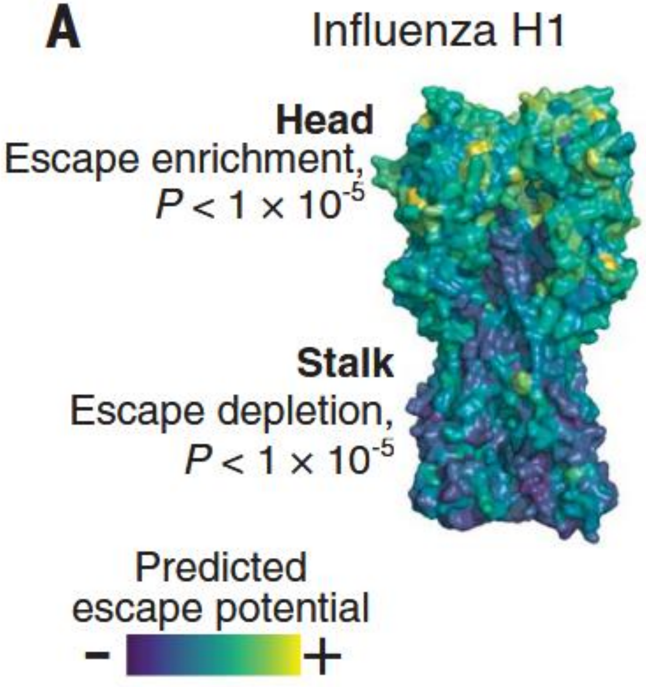
Escape Prediction Performance

Indeed, mutant sequences with high semantic change and high grammaticality are likely to promote viral escape



Structural Localization of Escape

Visualizing the escape potential across the viral protein structures reveals interesting patterns consistent with experimental findings



Outline

1. Background

1. Viruses & Viral Escape
2. Experimental Escape Modeling
3. Computational Escape Modeling
4. Analogy to Language Modeling

2. Methodology

1. Overview of Approach
2. Viral Protein Language Model
3. Viral Escape Prediction
4. Experiments & Implementation

3. Results

1. Semantic Embedding Clustering
2. Semantics & Grammaticality
3. Escape Prediction Performance
4. Structural Localization of Escape

4. Discussion

1. Key Findings & Takeaways
2. Strengths & Limitations
3. Future Directions
4. Concluding Remarks

Key Findings & Takeaways

1. There exists a strong, exploitable analogy between natural language and viral amino acid sequences

2. Mutations that significantly change viral “semantics” but preserve viral “grammaticality” cause viral escape

3. Estimates of viral “semantics” and “grammaticality” can be combined in a simple formula to assess escape likelihood

Strengths & Limitations

The strengths and limitations reflect the standard tensions between computational and experimental methods in biomedicine

Strengths

Highly scalable

Self-supervised

Strong conceptual contribution, with each component empirically validated

Readily handles combinatorial mutations

Limitations

Correlational and suggestive rather than causal and definitive

Only considers substitutions; does not natively consider insertions or deletions

Requires experimentally-derived amino acid sequences for training

Future Directions

1. Apply pipeline & CSCS to other instances of natural selection
2. Validate in the context of vaccine design
3. Incorporate information about post-translational changes
4. Explore alternative sequence modeling architectures

Concluding Remarks

Relating open problems in one domain to solved problems in another is a viable strategy for making scientific discoveries & progress

Language models learn surprisingly rich representations that can be harnessed in creative ways to solve hard downstream problems

Selecting the right variables greatly simplifies the task of modeling

References

1. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* 371, 284–288 (2021).
2. Taylor, M. W. What Is a Virus? in Viruses and Man: A History of Interactions (ed. Taylor, M. W.) 23–40 (Springer International Publishing, Cham, 2014). doi:10.1007/978-3-319-07758-1_2.
3. Viruses: What They Are & How They Work. Cleveland Clinic https://my.clevelandclinic.org/health/body/24861-virus.
4. Segre, J. Virus. National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Virus (2026).
5. Vidyasagar, A. What are viruses? | Live Science. https://www.livescience.com/53272-what-is-a-virus.html (2022).
6. Walsh, D. & Mohr, I. Viral subversion of the host protein synthesis machinery. Nat Rev Microbiol 9, 860–875 (2011).
7. Louten, J. Virus Replication. Essential Human Virology 49–70 (2016) doi:10.1016/B978-0-12-800947-5.00004-1.
8. The Viral Life Cycle | Microbiology. *Lumen Learning* https://courses.lumenlearning.com/suny-microbiology/chapter/the-viral-life-cycle/.
9. Lucas, M., Karrer, U., Lucas, A. & Klenerman, P. Viral escape mechanisms – escapology taught by viruses. Int J Exp Pathol 82, 269–286 (2001).
10. Koyama, S., Ishii, K. J., Coban, C. & Akira, S. Innate immune response to viral infection. Cytokine 43, 336–341 (2008).
11. Fig. 1: The life cycle of influenza virus and its antibody targets. | Nature Reviews Immunology. https://www.nature.com/articles/s41577-019-0143-6/figures/1.
12. Lee, J. M. et al. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. eLife 8, e49324 (2019).
13. Allman, B. E., Vieira, L., Diaz, D. J. & Wilke, C. O. A systematic evaluation of the language-of-viral-escape model using multiple machine learning frameworks. J R Soc Interface 22, 20240598.
14. Wang, T. T. & Palese, P. Universal epitopes of influenza virus hemagglutinins? Nat Struct Mol Biol 16, 233–234 (2009).
15. PDB101: Molecule of the Month: HIV Envelope Glycoprotein. RCSB: PDB-101 http://pdb101.rcsb.org/motm/169.
16. Florindo, H. F. et al. Immune-mediated approaches against COVID-19. Nat. Nanotechnol. 15, 630–645 (2020).
17. Louvain — scikit-network 0.33.0 documentation. https://scikit-network.readthedocs.io/en/latest/tutorials/clustering/louvain.html.