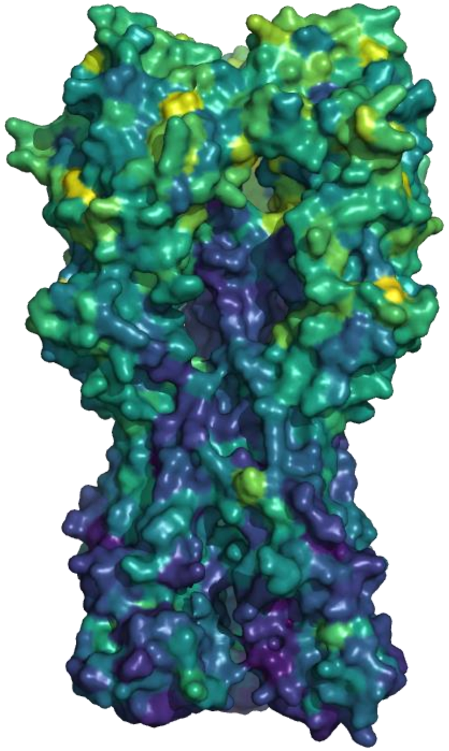
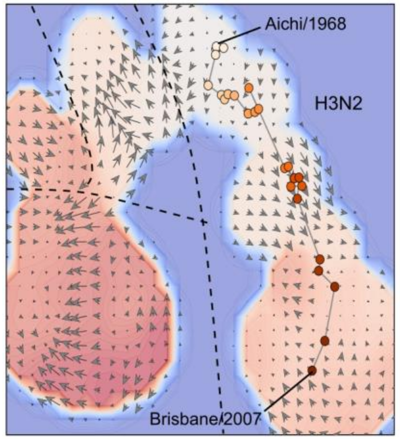
Predicting pathogen evolution with neural language models

Brian Hie

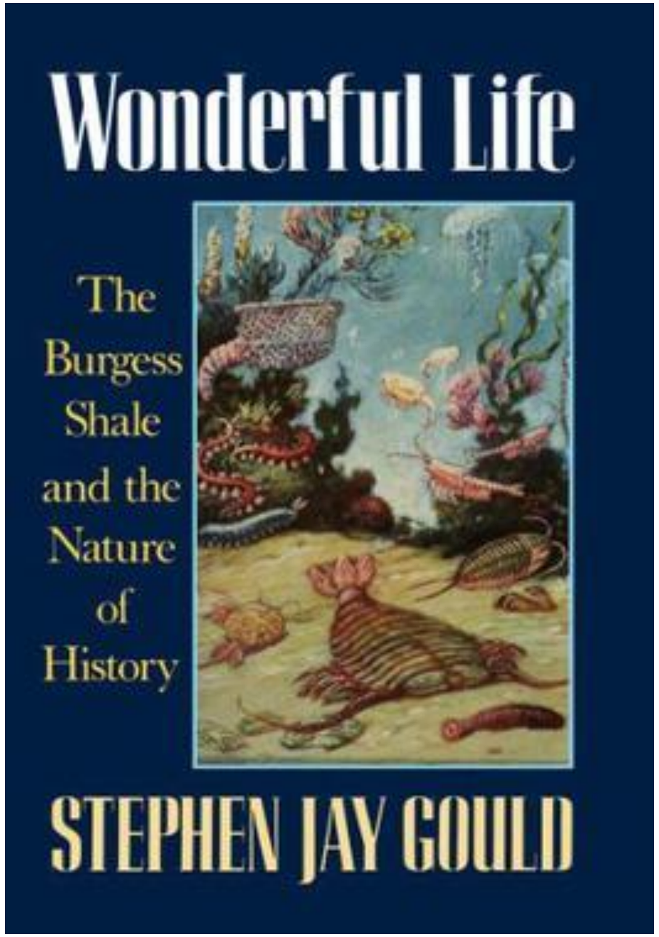
with Bonnie Berger, Bryan Bryson, Peter Kim, Kevin Yang, and Ellen Zhong

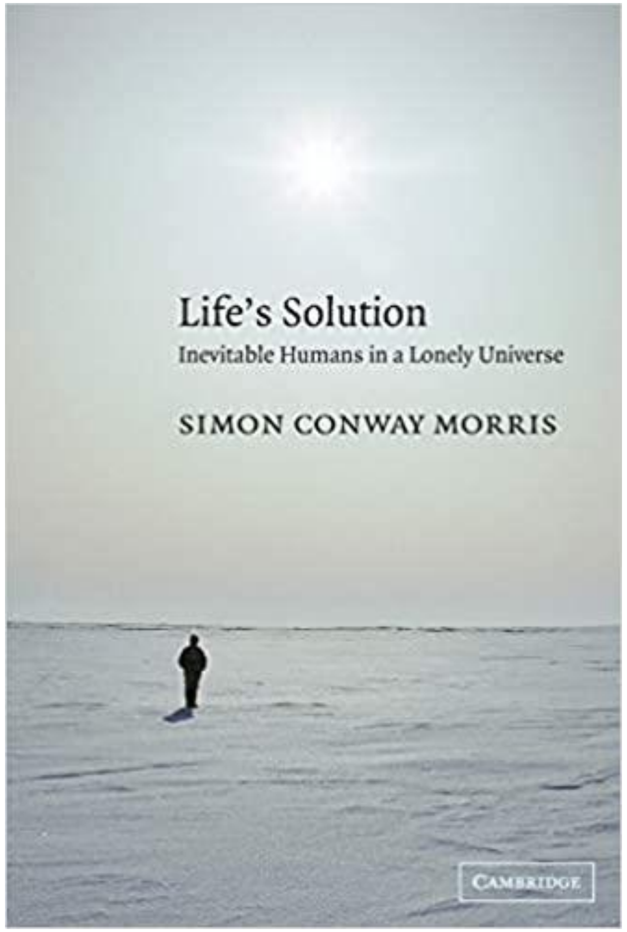
October 28, 2021





How predictable is evolution?





1. Learning the language of viral evolution and escape

2. Recovering evolutionary dynamics with “evolutionary velocity”

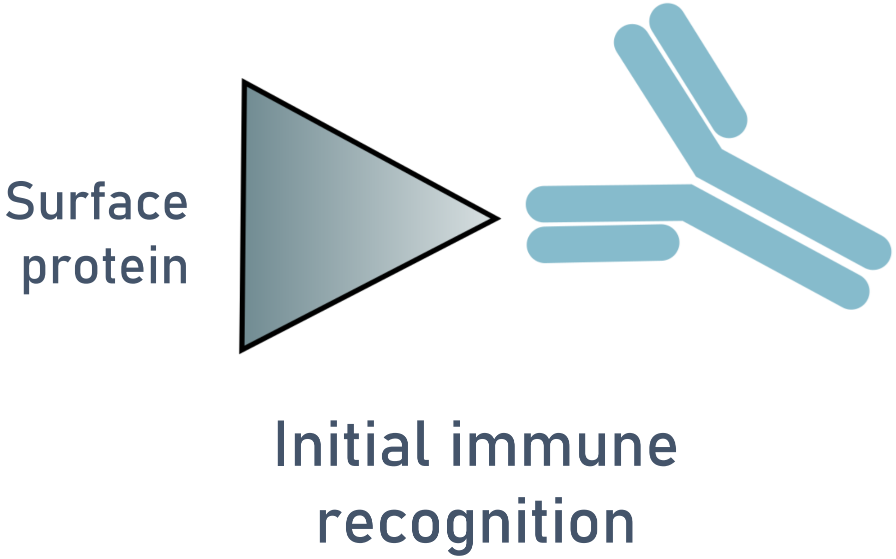
3. Looking forward

1. Learning the language of viral evolution and escape

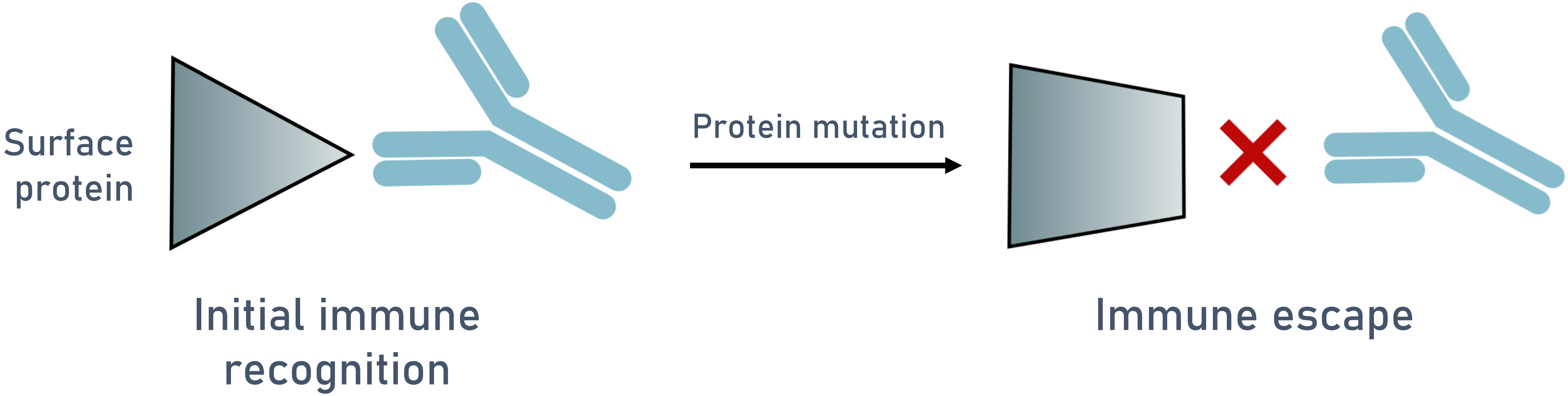
2. Recovering evolutionary dynamics with “evolutionary velocity”

3. Looking forward

Viral escape

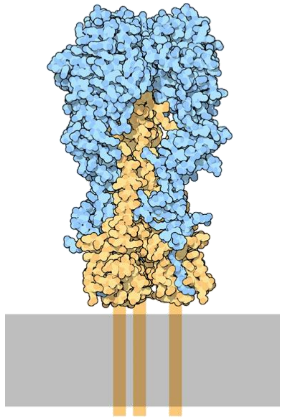


Viral escape

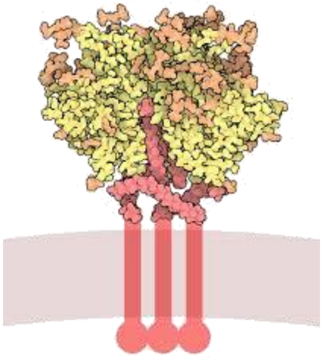


Viral escape is a big problem

Influenza A HA



HIV Env



Influenza

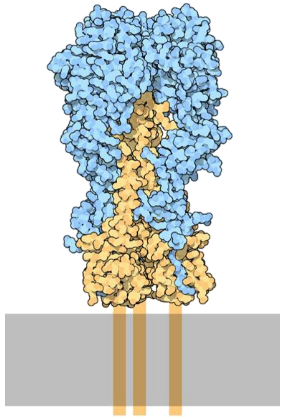
- 250K-600K deaths a year
- Yearly vaccine that is 20-50% effective

AIDS

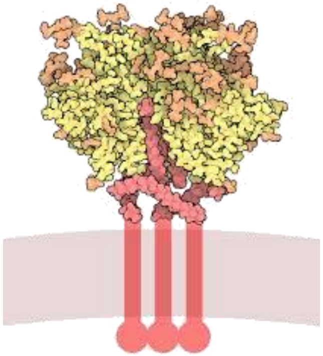
- 700K-1.2M deaths a year
- No effective vaccine

Viral escape is a big problem

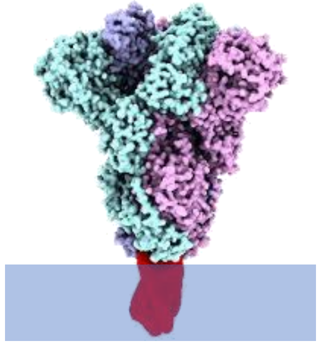
Influenza A HA



HIV Env



SARS-CoV-2 Spike



Influenza
- 250K-600K deaths a year
- Yearly vaccine that is 20-50% effective

AIDS
- 700K-1.2M deaths a year
- No effective vaccine

COVID-19
- 4.9M+ deaths
- Questions about durability of vaccine protection

Small changes can have big semantic effects

Small changes can have big semantic effects

The boy pats the dog.

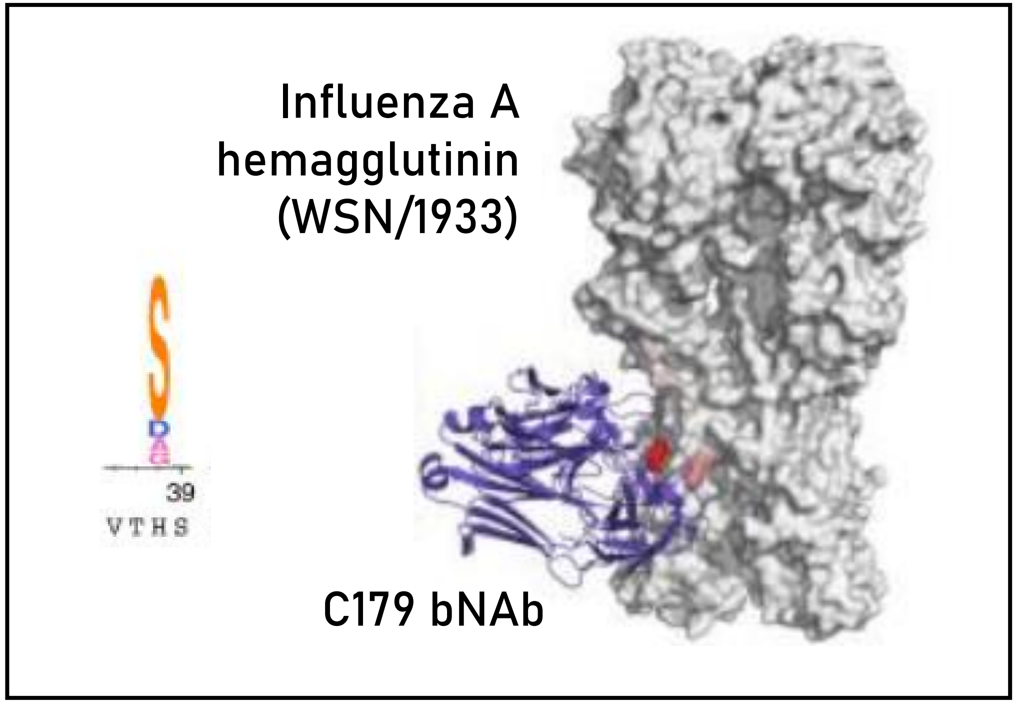
Small changes can have big semantic effects

The boy pats the dog.

vs

The boy eats the dog.

Single residue change enables viral escape

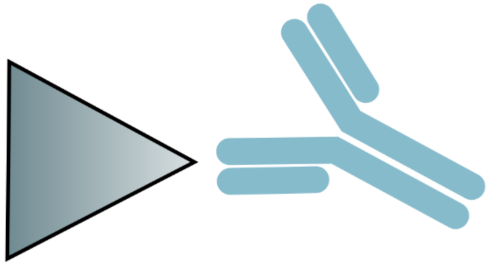


From Doud, Lee, and Bloom. *Nat. Comm.* (2018)

H → S mutation means C179 no longer binds

Connection to biology

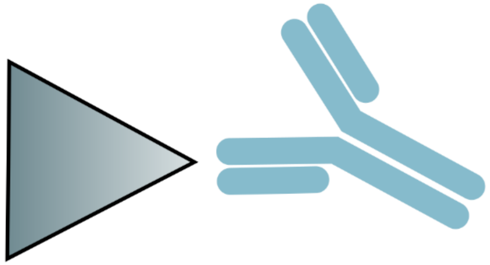
The language of viral escape



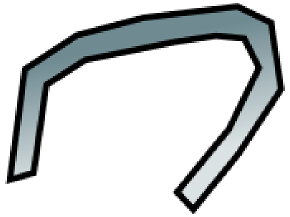
The boy pats the dog.

Connection to biology

The language of viral escape



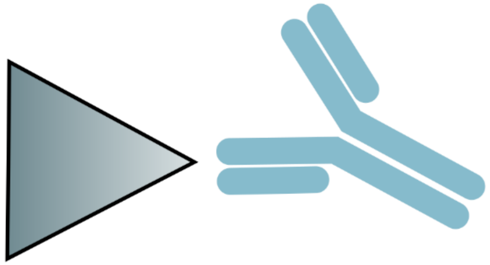
The boy pats the dog.



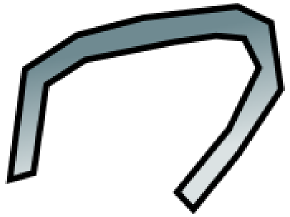
The boy patx the dog.

Connection to biology

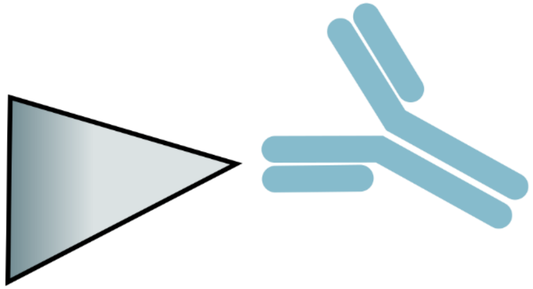
The language of viral escape



The boy pats the dog.



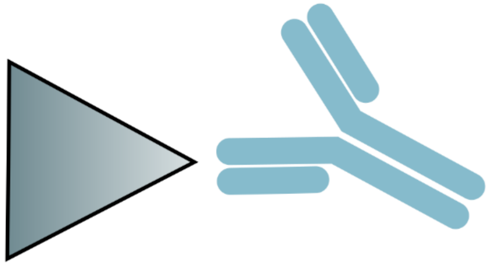
The boy patx the dog.



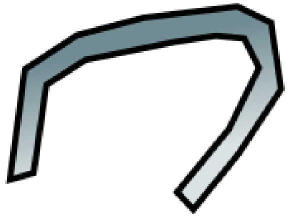
The boy pets the dog.

Connection to biology

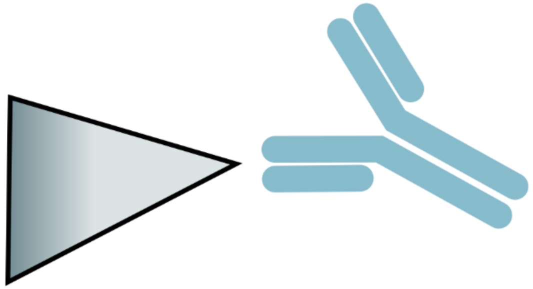
The language of viral escape



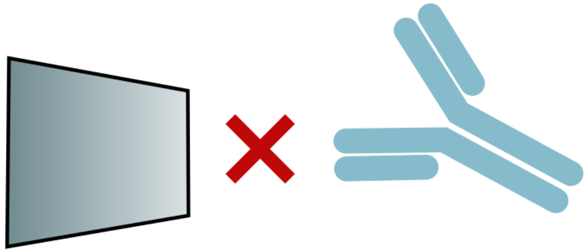
The boy pats the dog.



The boy patx the dog.



The boy pets the dog.



The boy eats the dog.

Our approach

Constrained semantic change search (CSCS)

- You're given a sequence of tokens from some language

- Goal: Find the single token change that:

1. Induces the largest semantic change

2. Is constrained by the rules/grammar of that language

Our approach

Constrained semantic change search (CSCS)

- You’re given a sequence of tokens from some language

- Goal: Find the single token

1. Induces the largest sema
2. Is constrained by the rules

For example:

A sequence of words from an English sentence

Or

A sequence of amino acids from a viral protein

Some real example changes...

Original headline:

australian dead in bali

Some real example changes...

Original headline:
australian dead in bali

Semantically closest:
aussie dead in bali

Some real example changes...

Original headline:
australian dead in bali

Semantically closest:
aussie dead in bali

CSCS proposed change:
australian ballet in bali

Some real example changes...

Original headline:
blast off of apollo 8

Semantically closest:
blast off of apollo 13

CSCS proposed change:
blast victims of apollo 8

Some real example changes...

Original headline:
excuse me you left a gorilla suit on the bus

Semantically closest:
excuse me we left a gorilla suit on the bus

CSCS proposed change:
excuse me you left a gorilla killer on the bus

Some real example changes...

Original headline:
winegrowers revel in good season

Semantically closest:
winegrowers revel in strong season

CSCS proposed change:
winegrowers revel in flu season

Some real example changes...

Original headline:
winegrowers revel in good season

Semantically closest:
winegrowers revel in strong season

CSCS proposed change:
winegrowers revel in flu season

Saving up my boogers for ten years and then building the world's largest booger.

Family Edition

My name is Peter Parker. I was bitten by a radioactive spider, and now I'm ________

Family Edition

Getting stuck in the toilet.

Family Edition

Thanks for watching! If you want to see more vids of ________, smash that subscribe.

Family Edition

Our approach

A computational language model

On natural language sequences:

The American president _____________ to Japan yesterday.

$p(\text{"went"}\ ) = 0.5$
$p(\text{"traveled"}) = 0.2$
$p(\text{"absconded"}) = 0.05$
$\vdots$
$p(\text{"xylophone"}) = 0$

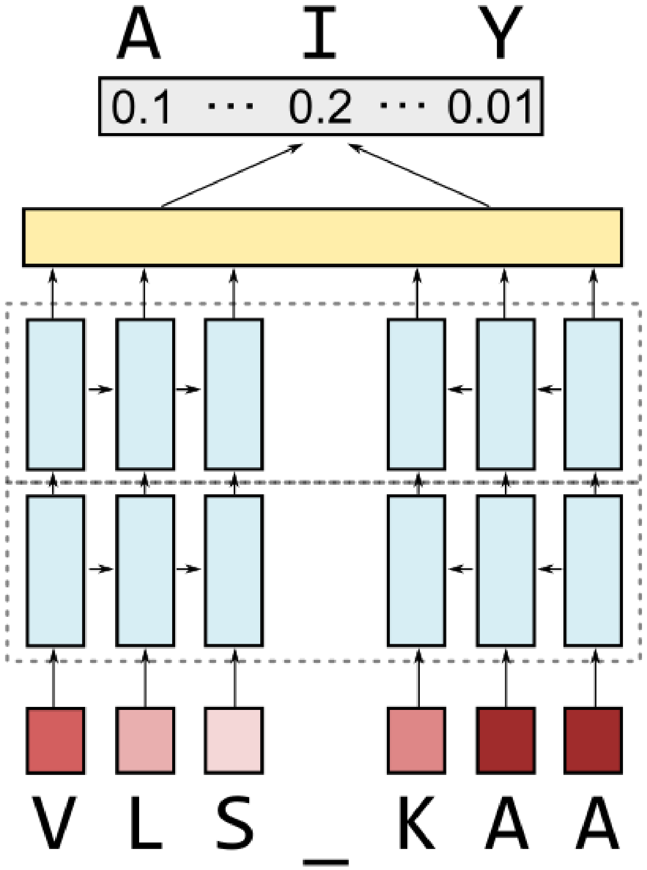
Our approach

A computational language model

On natural language sequences:

The American president _______________
to Japan yesterday.

$p(\text{"went"}) = 0.5$
$p(\text{"traveled"}) = 0.2$
$p(\text{"absconded"}) = 0.05$
$\vdots$
$p(\text{"xylophone"}) = 0$



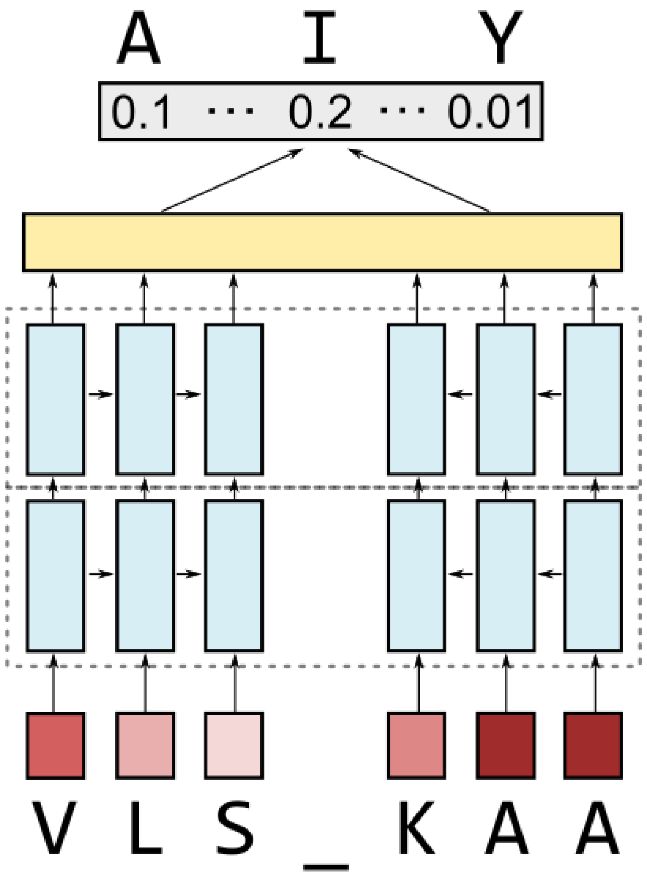
Our approach

A computational language model

On natural language sequences:

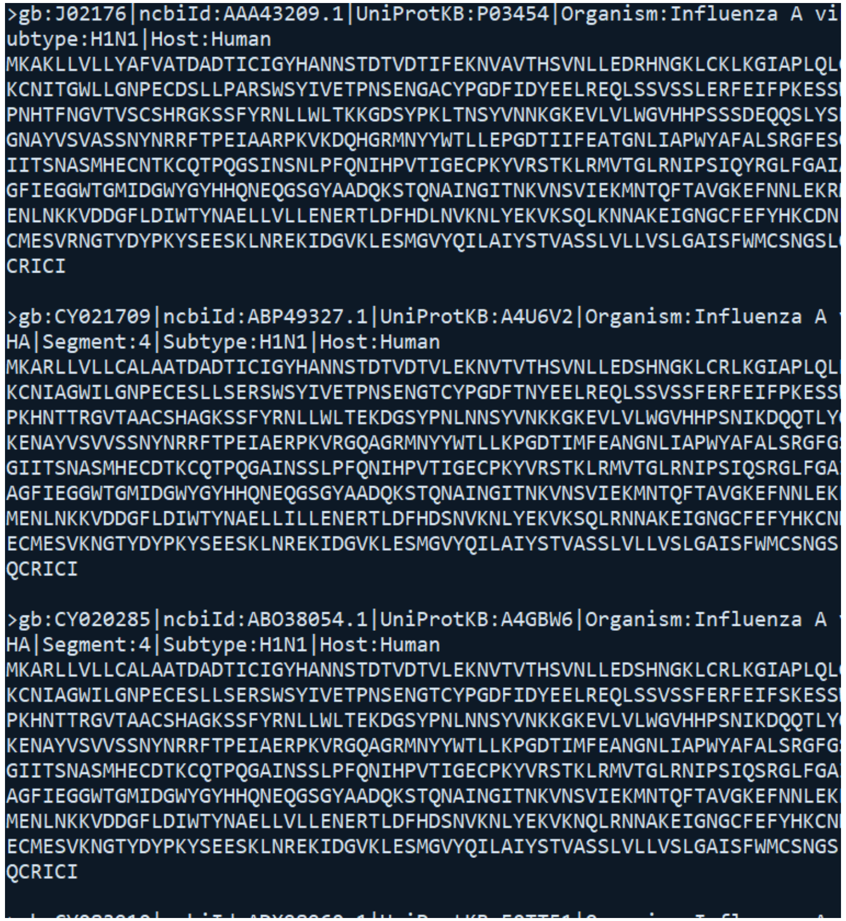
The American president ______________
to Japan yesterday.

$p(\text{"went"}) = 0.5$
$p(\text{"traveled"}) = 0.2$
$p(\text{"absconded"}) = 0.05$
$\vdots$
$p(\text{"xylophone"}) = 0$



Trained on thousands of sequences (or more)!

Train on viral protein sequence corpus



```
>gb:J02176|ncbiId:AAA43209.1|UniProtKB:P03454|Organism:Influenza A vi
ubtype:H1N1|Host:Human
MKAKLLVLLYAFVATDADTICIGYHANNSTDTVDTIFEKNVAVTHSVNLLEDRHNGKLCKLKGIAPLQL
KCNITGWLLGNPECDSLLPARSWSYIVETPNSENGACYPGDFIDYEELREQLSSVSSLERFEIFPKESS
PNHTFNGVTVSCSHRGKSSFYRNLLWLTKKGDSYPKLTNSYVNNKGKEVLVLWGVHHPSSSDEQQSLYS
GNAYVSVASSNYNRRFTPEIAARPKVKDQHGRMNYYWTLLEPGDTIIFEATGNLIAPWYAFALSRGFES
IITSNASMHECNTKCQTPQGSINSNLPFQNIHPVTIGECPKYVRSTKLRMVTGLRNIPSIQYRGLFGAI
GFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAINGITNKVNSVIEKMNTQFTAVGKEFNNLEKR
ENLNKKVDDGFLDIWTYNAELLVLLENERTLDFHDLNVKNLYEKVKSQLKNNAKEIGNGCFEFYHKCDN
CMESVRNGTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAISFWMCSNGSL
CRICI

>gb:CY021709|ncbiId:ABP49327.1|UniProtKB:A4U6V2|Organism:Influenza A
HA|Segment:4|Subtype:H1N1|Host:Human
MKARLLVLLCALAATDADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLKGIAPLQL
KCNIAGWILGNPECESLLSERSWSYIVETPNSENGTCYPGDFTNYEELREQLSSVSSFERFEIFPKESS
PKHNTTRGVTAACSHAGKSSFYRNLLWLTEKDGSYPNLNNSYVNKKGKEVLVLWGVHHPSNIKDQQTLY
KENAYVSVVSSNYNRRFTPEIAERPKVRGQAGRMNYYWTLLKPGDTIMFEANGNLIAPWYAFALSRGFG
GIITSNASMHECDTKCQTPQGAINSSLPFQNIHPVTIGECPKYVRSTKLRMVTGLRNIPSIQSRGLFGA
AGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAINGITNKVNSVIEKMNTQFTAVGKEFNNLEK
MENLNKKVDDGFLDIWTYNAELLILLENERTLDFHDSNVKNLYEKVKSQLRNNAKEIGNGCFEFYHKCN
ECMESVKNGTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAISFWMCSNGS
QCRICI

>gb:CY020285|ncbiId:ABO38054.1|UniProtKB:A4GBW6|Organism:Influenza A
HA|Segment:4|Subtype:H1N1|Host:Human
MKARLLVLLCALAATDADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLKGIAPLQL
KCNIAGWILGNPECESLLSERSWSYIVETPNSENGTCYPGDFIDYEELREQLSSVSSFERFEIFSKESS
PKHNTTRGVTAACSHAGKSSFYRNLLWLTEKDGSYPNLNNSYVNKKGKEVLVLWGVHHPSNIKDQQTLY
KENAYVSVVSSNYNRRFTPEIAERPKVRGQAGRMNYYWTLLKPGDTIMFEANGNLIAPWYAFALSRGFG
GIITSNASMHECDTKCQTPQGAINSSLPFQNIHPVTIGECPKYVRSTKLRMVTGLRNIPSIQSRGLFGA
AGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAINGITNKVNSVIEKMNTQFTAVGKEFNNLEK
MENLNKKVDDGFLDIWTYNAELLVLLENERTLDFHDSNVKNLYEKVKNQLRNNAKEIGNGCFEFYHKCN
ECMESVKNGTYDYPKYSEESKLNREKIDGVKLESMGVYQILAIYSTVASSLVLLVSLGAISFWMCSNGS
QCRICI
```

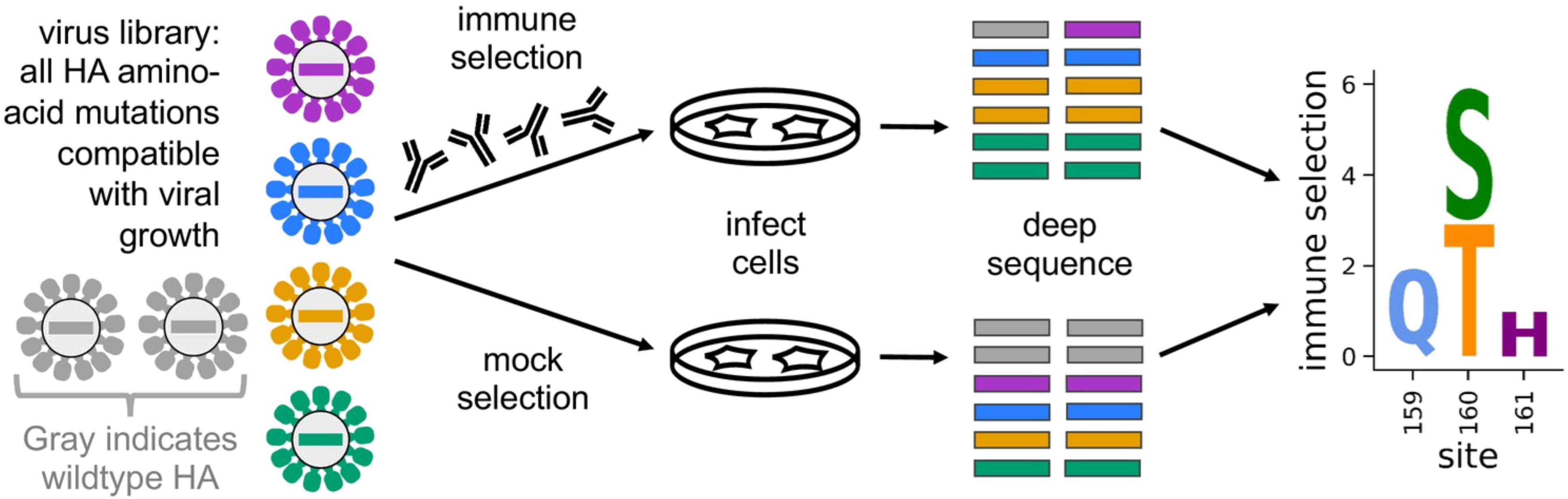
Influenza A HA
NIAID Influenza Research Database
(https://www.fludb.org)

HIV-1 Env
LANL HIV Database
(https://www.hiv.lanl.gov)

SARS-CoV-2 Spike
Virus Pathogen Database angd Analysis Resource
(https://www.viprbrc.org/)
GISAID
(https://www.gisaid.org/)
NCBI GenBank
(https://www.ncbi.nlm.nih.gov/sars-cov-2/)

Results

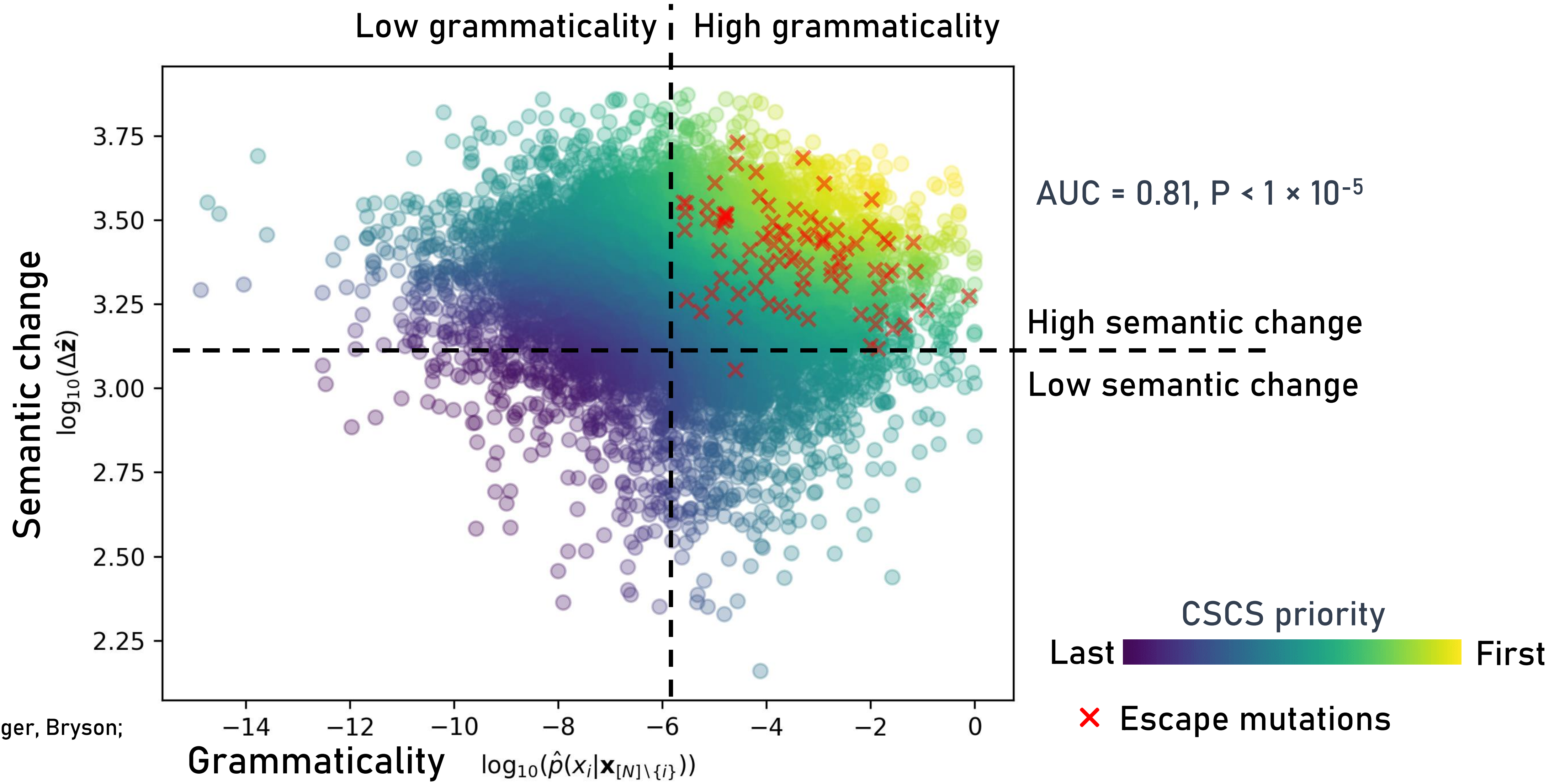
Putting it all together to predict viral escape



From Lee et al., *eLife* (2019)

Results

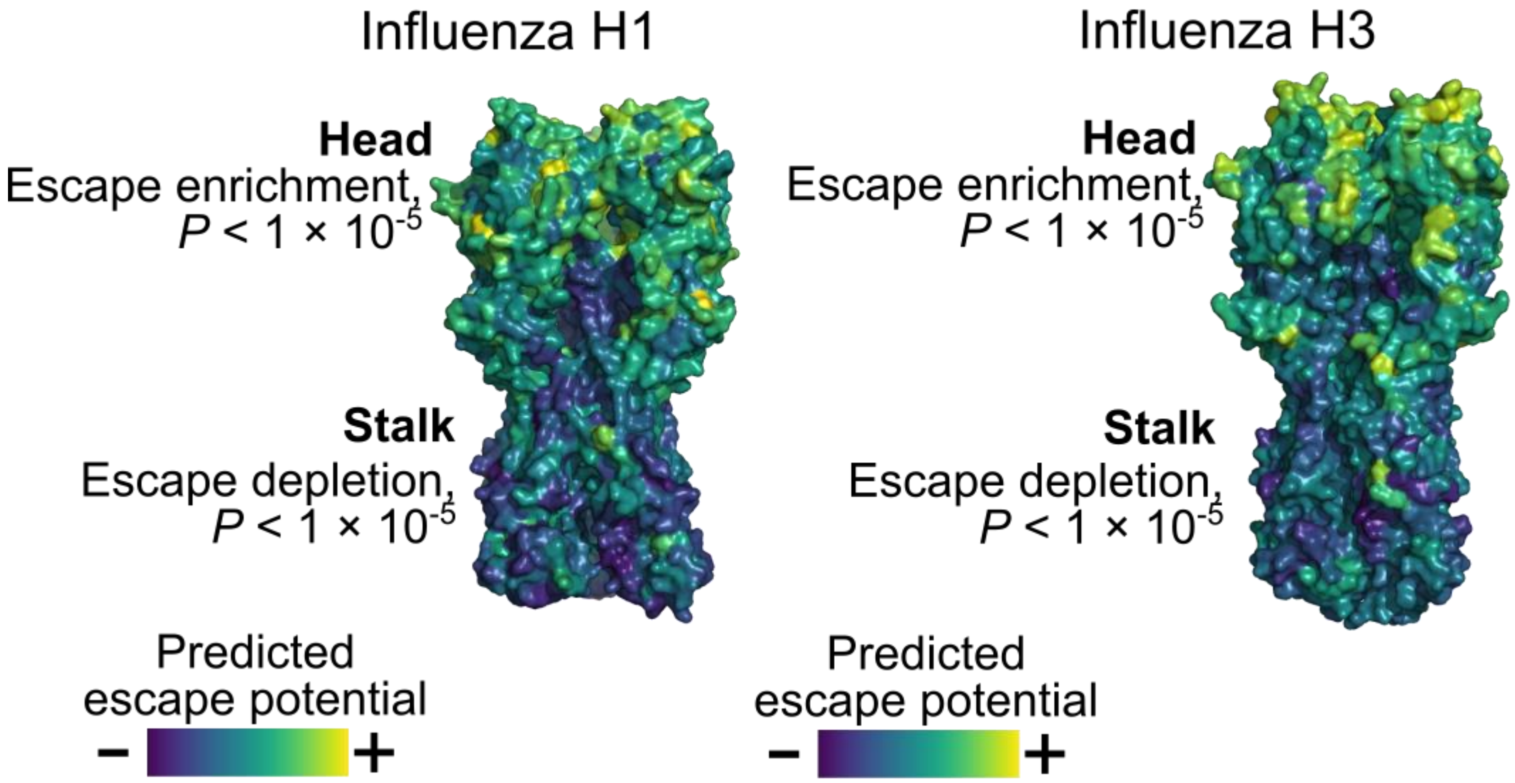
Unsupervised prediction of escape mutations



From Hie, Zhong, Berger, Bryson; *Science*, 2021

Results

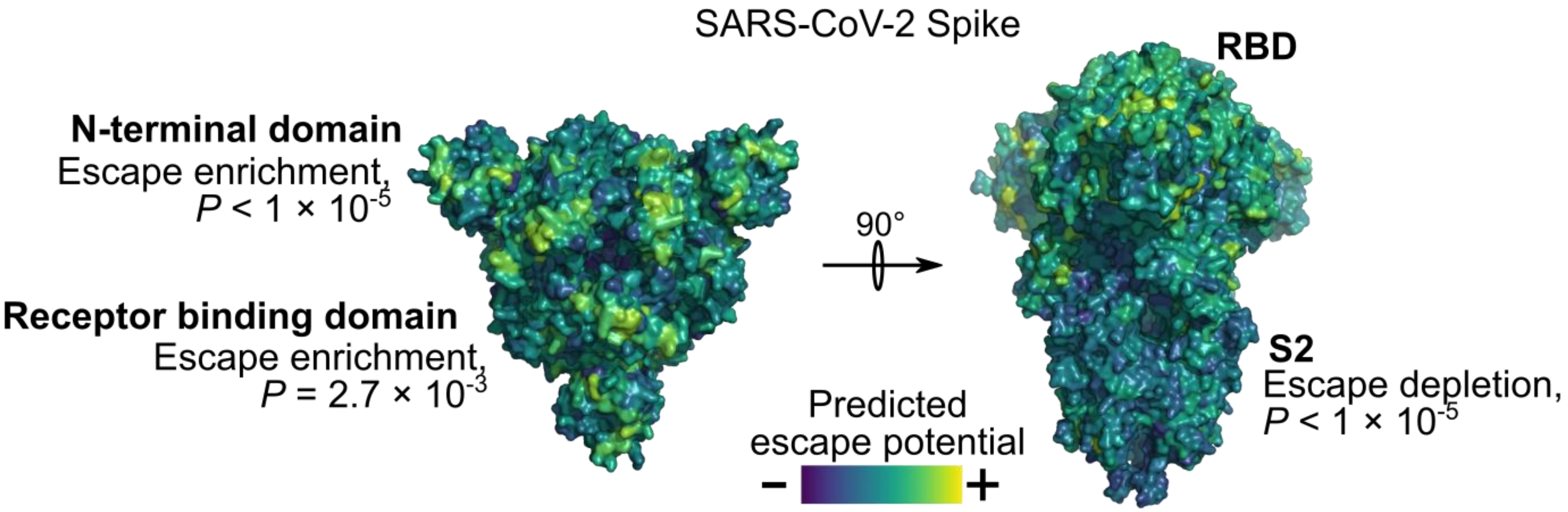
Enriched escape potential in HA head



From Hie, Zhong, Berger, Bryson; *Science*, 2021

Results

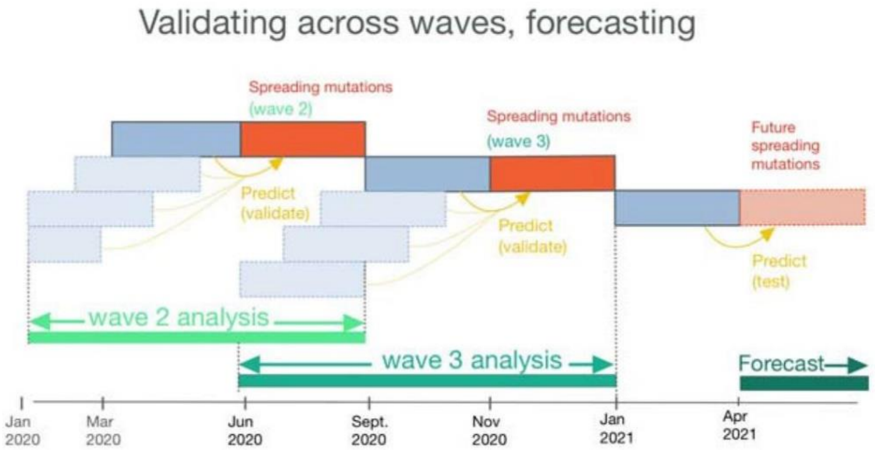
Similar patterns for CoV-2 S1 versus S2



From Hie, Zhong, Berger, Bryson; *Science*, 2021

Results

Language model predicts SARS-CoV-2 variants


Validating across waves, forecasting

Spike language model predicts mutations up to 4 months in advance with AUC of 0.8

Cyrus Maher

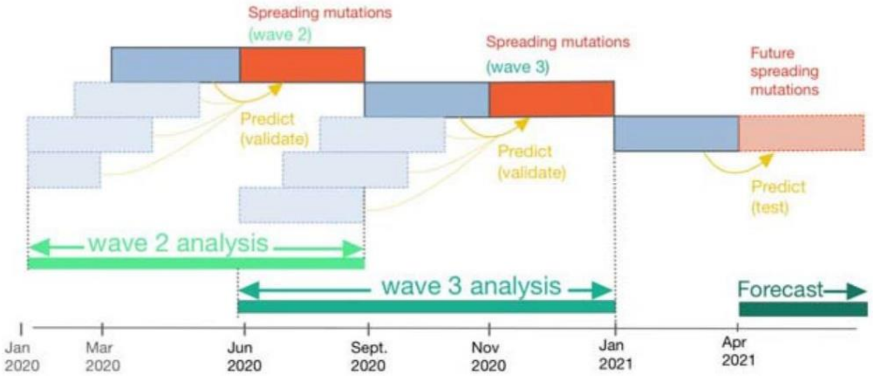
VIR

From Maher et al.; *medRxiv*, 2021

Results

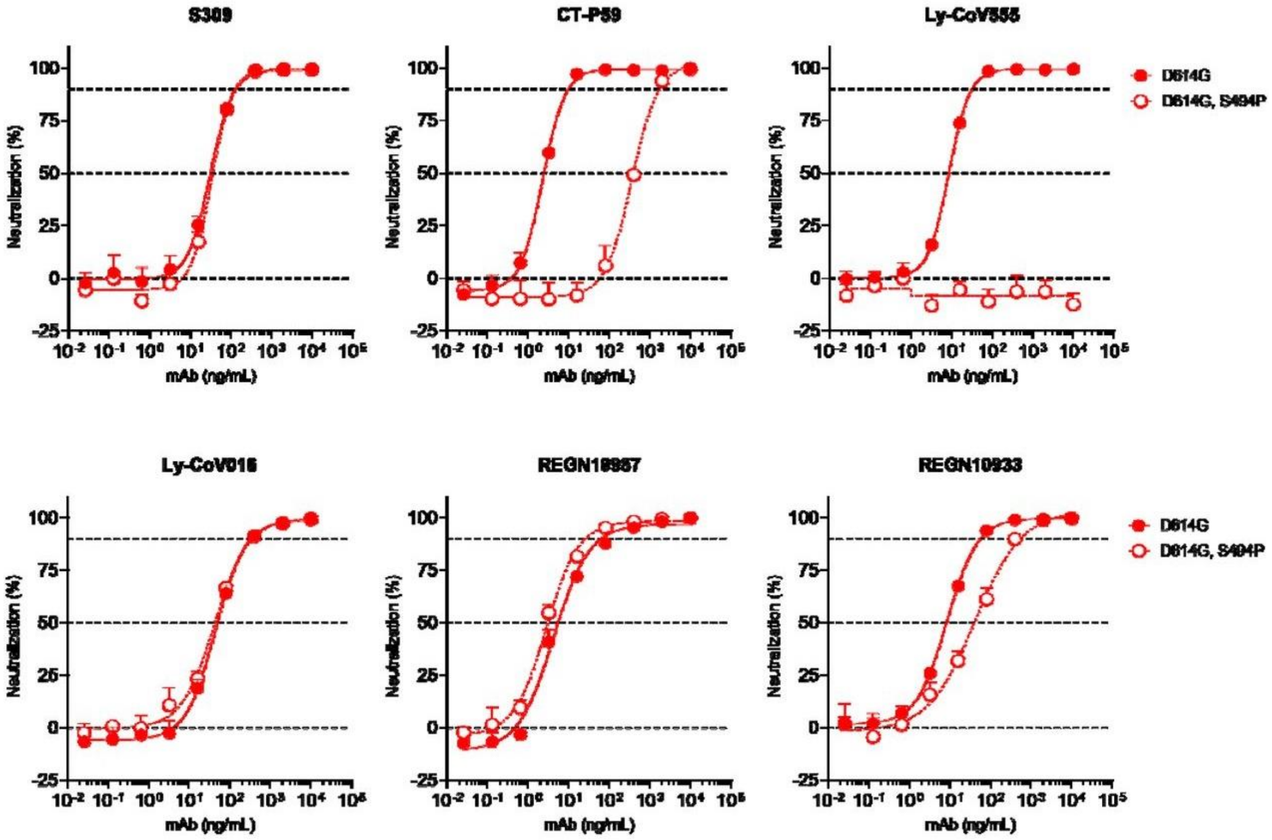
Language model predicts SARS-CoV-2 variants

Validating across waves, forecasting



Spike language model predicts mutations up to 4 months in advance with AUC of 0.8

Cyrus Maher



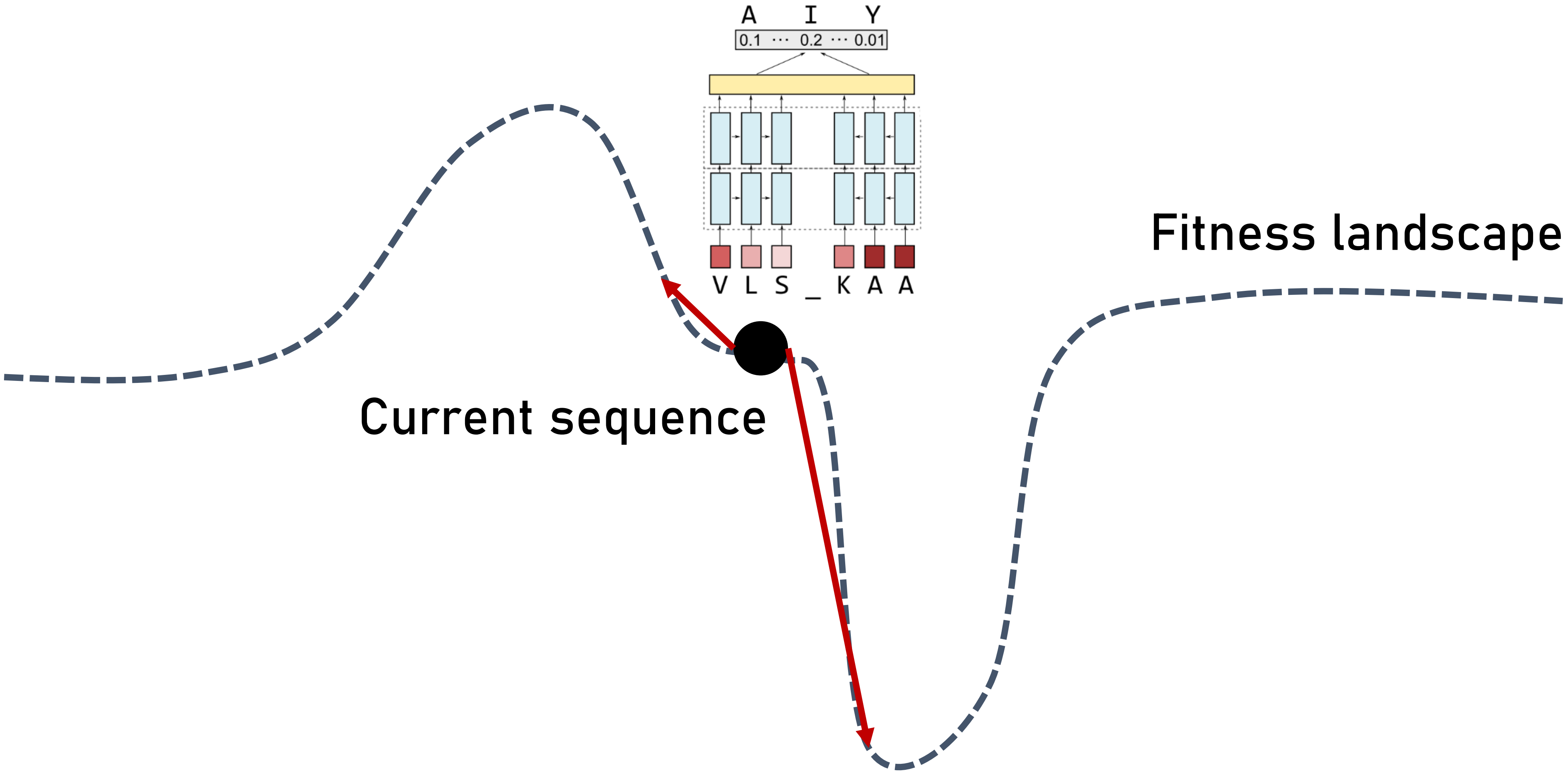
From Maher et al.; *medRxiv*, 2021

1. Learning the language of viral evolution and escape

2. **Recovering evolutionary dynamics with “evolutionary velocity”**

3. Looking forward

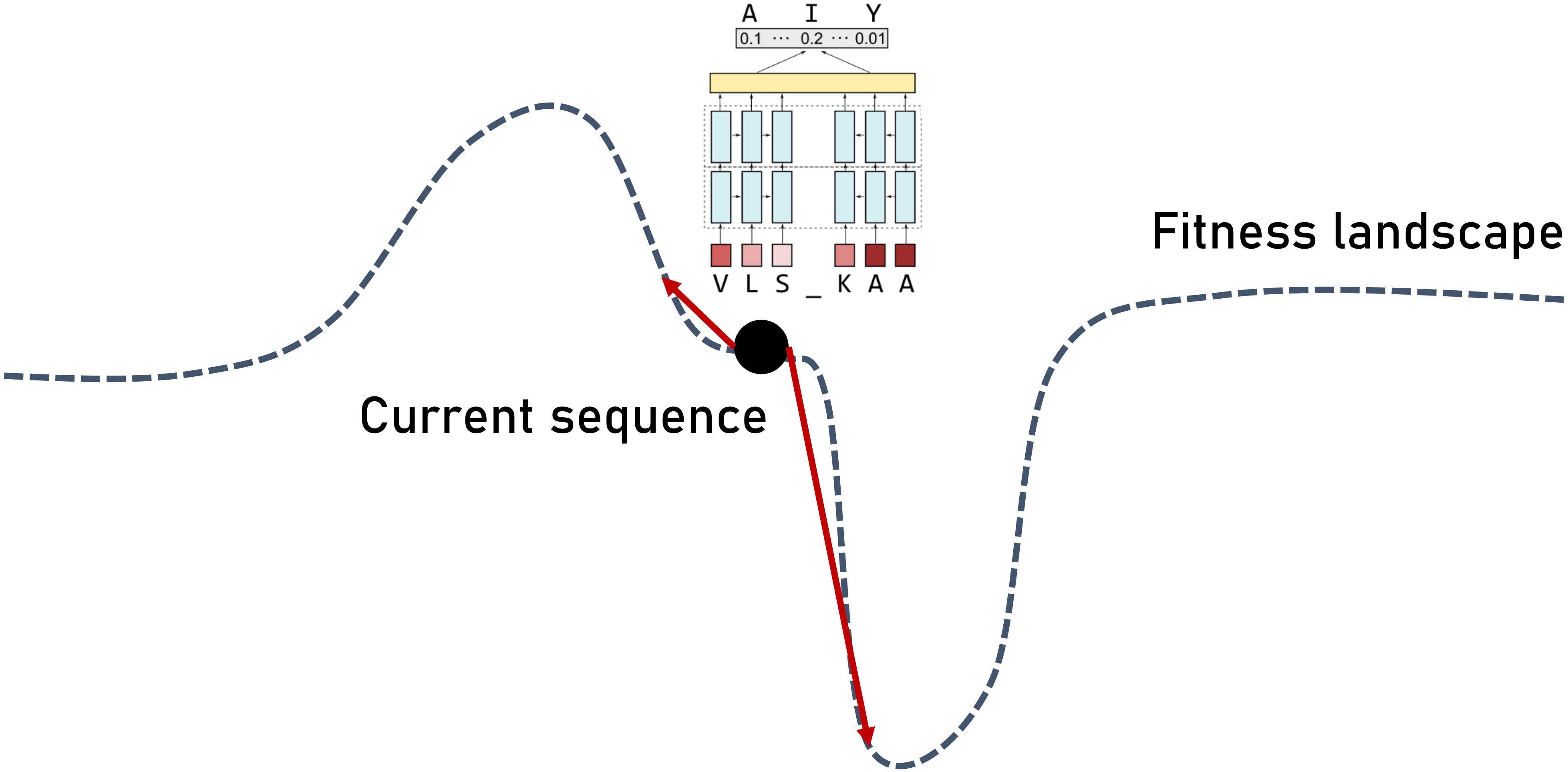
We can predict local evolution with LMs



Fitness landscape

Current sequence

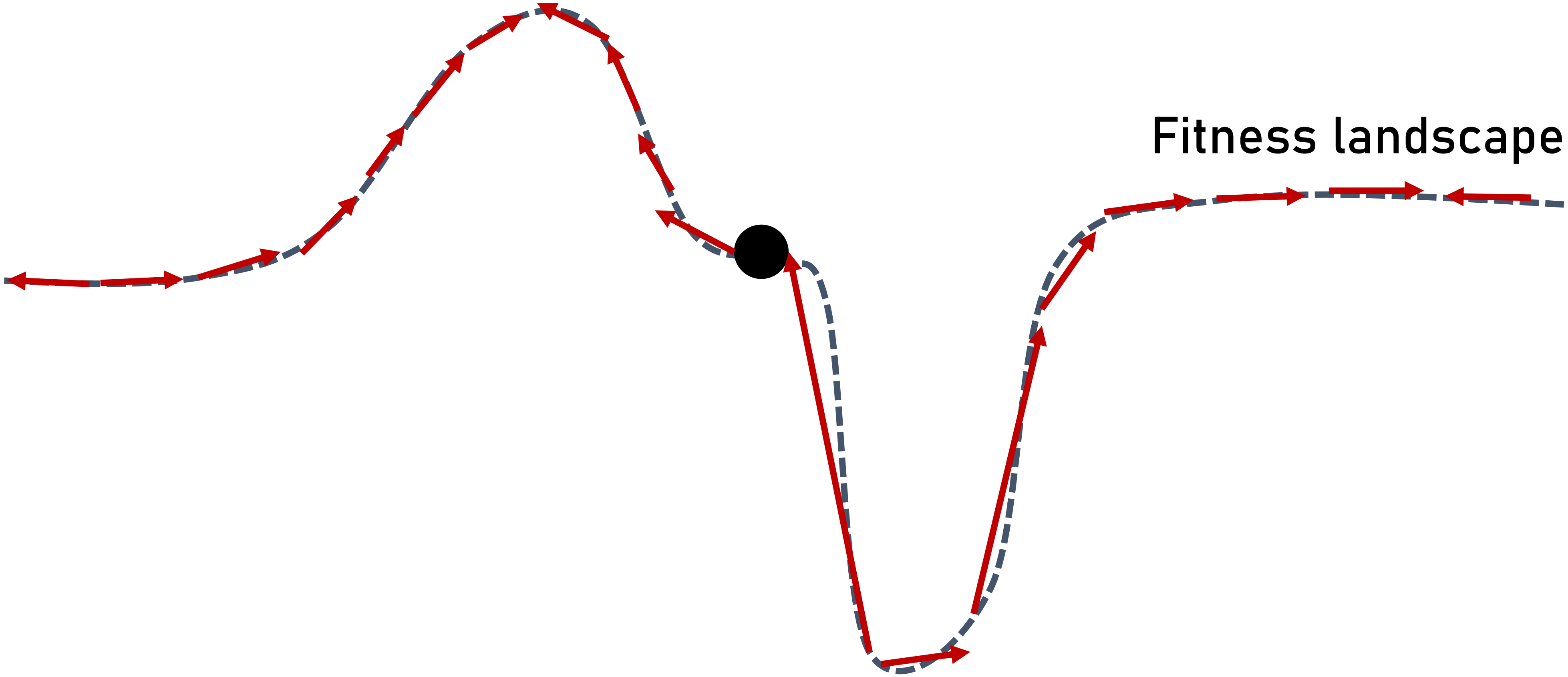
We can predict local evolution with LMs



Big open question: What about > 1-residue mutations?

Idea

Understand global patterns using local predictions

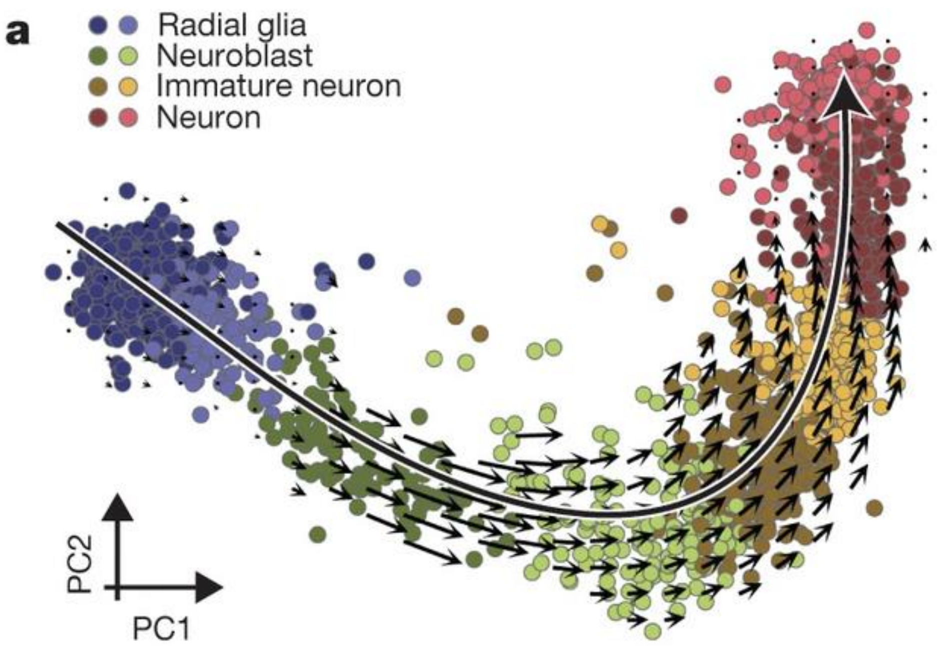


Idea

Understand global patterns using local predictions

"RNA velocity"

(La Manno et al, *Nature*, 2018)

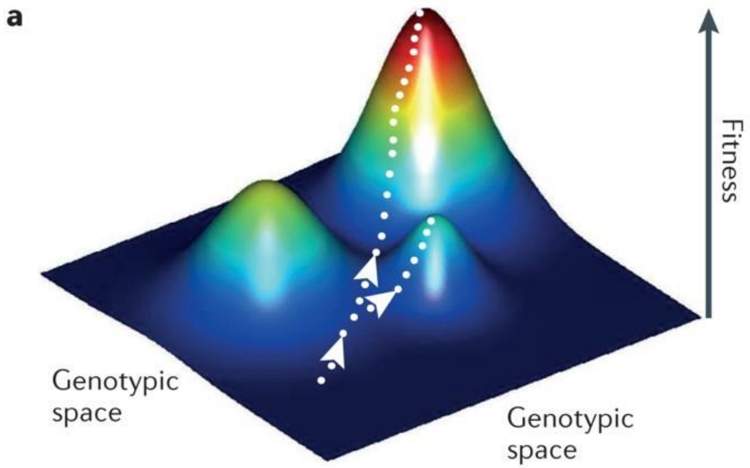


Idea

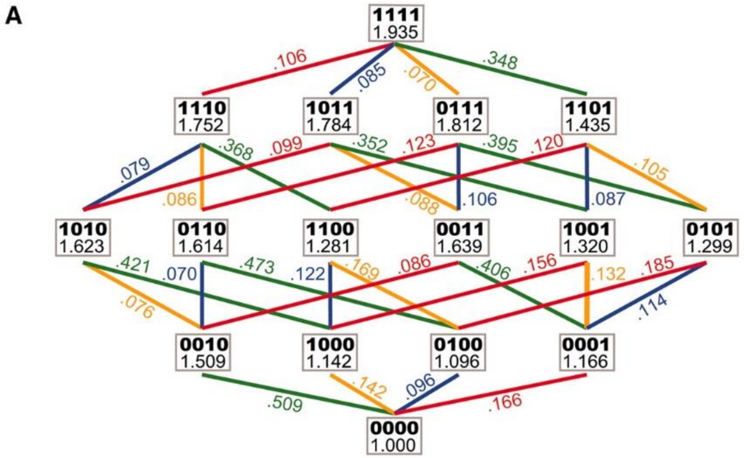
Understand global patterns using local predictions

“Fitness landscape”

(Wright, *Int. Conf. Genetics*, 1932)

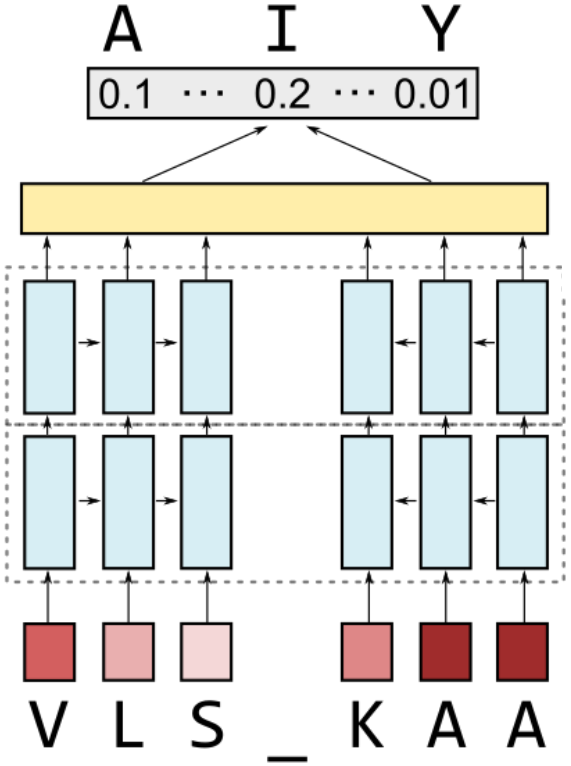


Visser and Krug. *Nat. Rev. Genetics* (2014)



Chou et al. *Science* (2011)

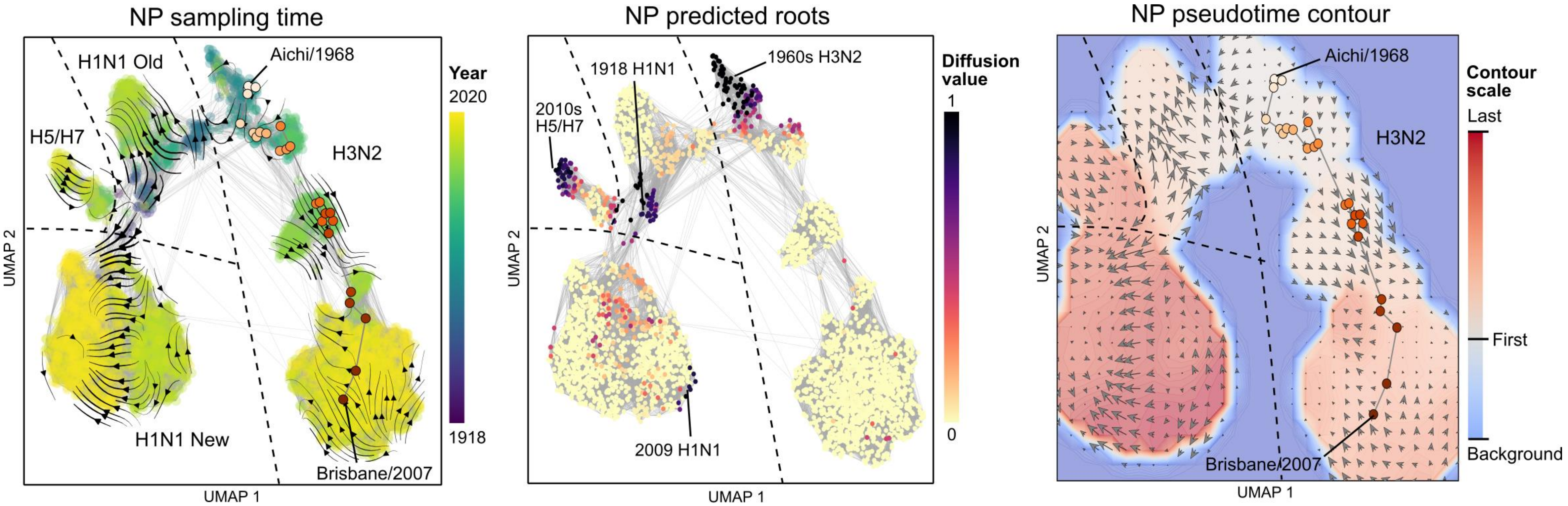
“Universal” protein language model



ESM-1b by Rives et al. (*PNAS*, 2021):

- Trained on 3 million sequences from UniRef50
- Model has 650 million parameters

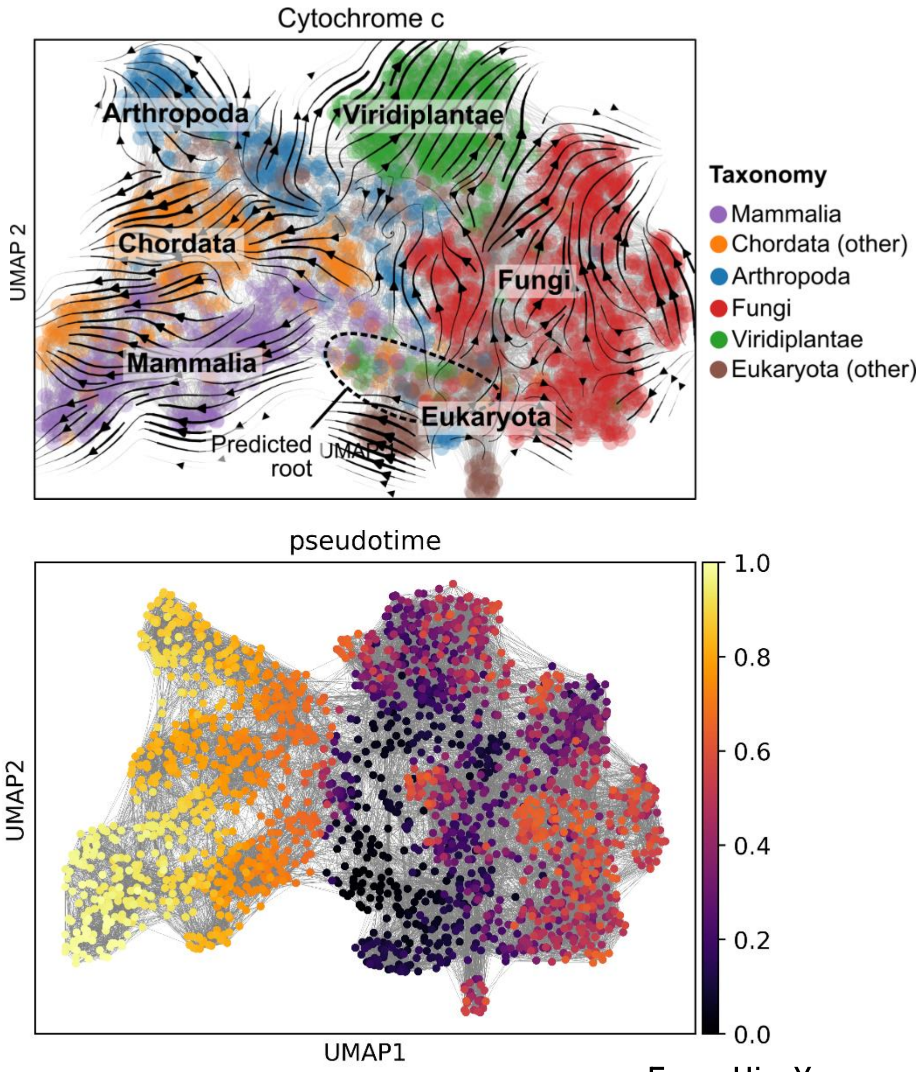
Velocity of influenza evolution

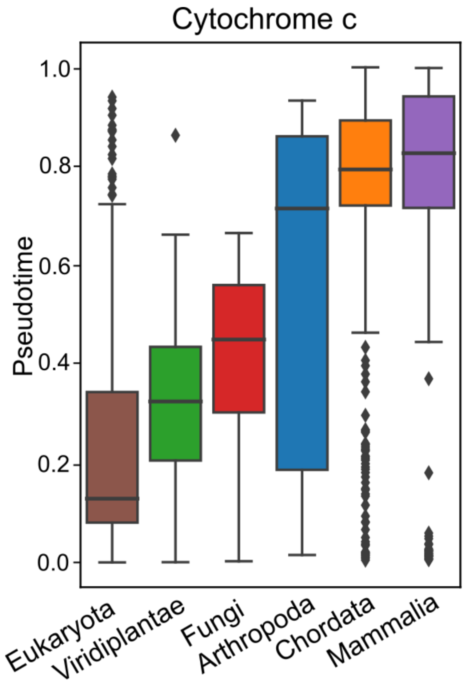


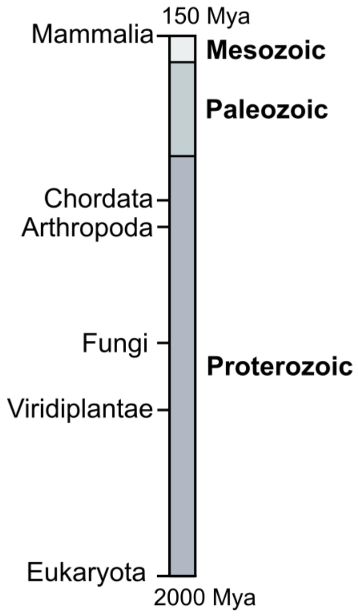
Temporal Spearman r = 0.49, P < 1e-308

From Hie, Yang, and Kim; *bioRxiv*, 2021

Velocity of cytochrome c evolution

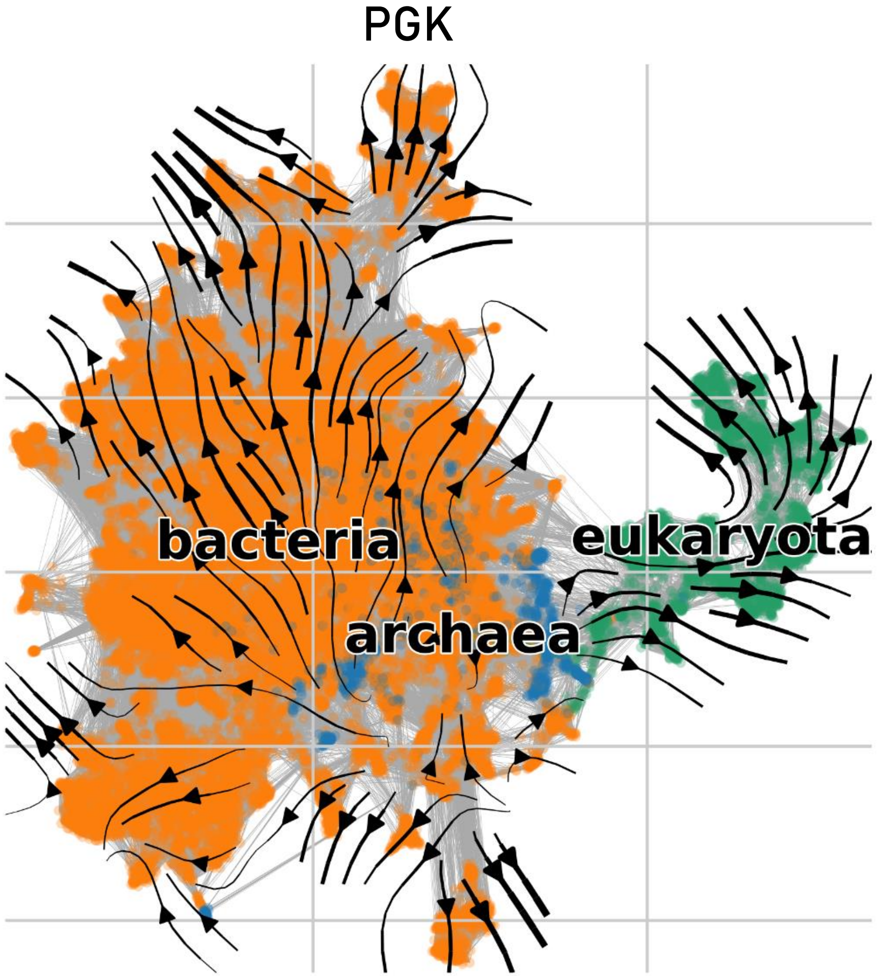


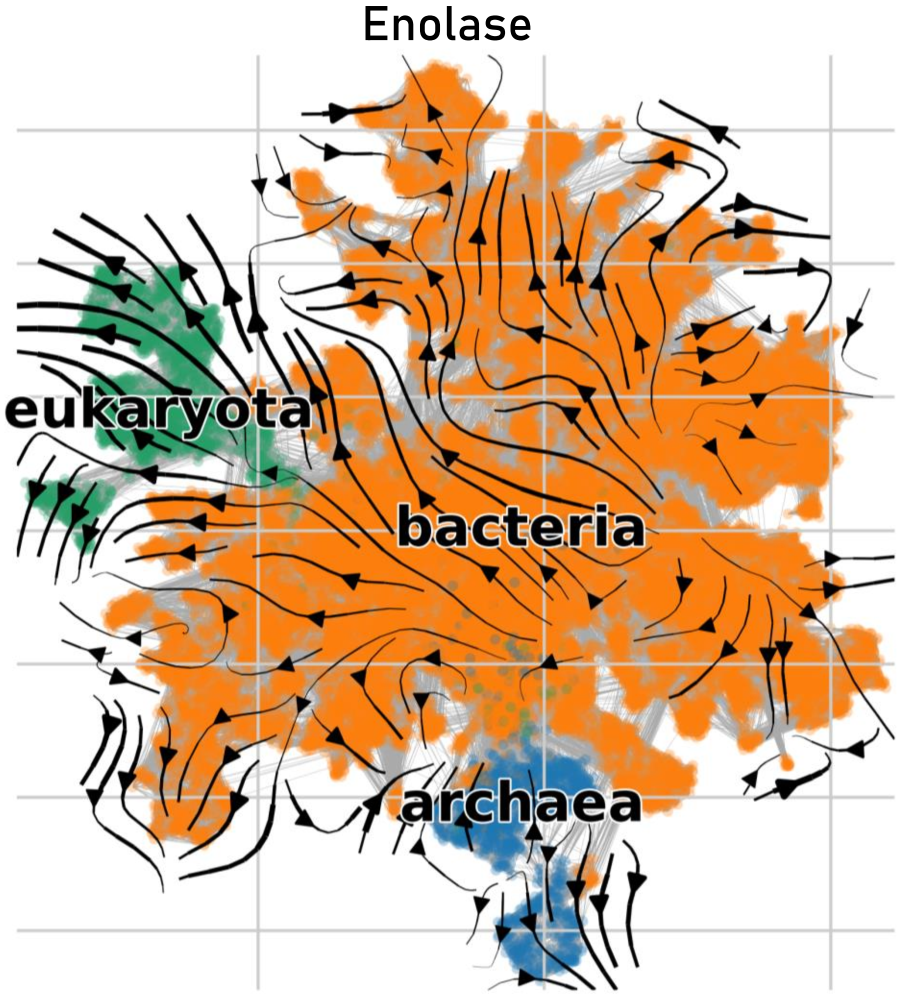




From Hie, Yang, and Kim; *bioRxiv*, 2021

Velocity of ancient evolution





From Hie, Yang, and Kim; *bioRxiv*, 2021

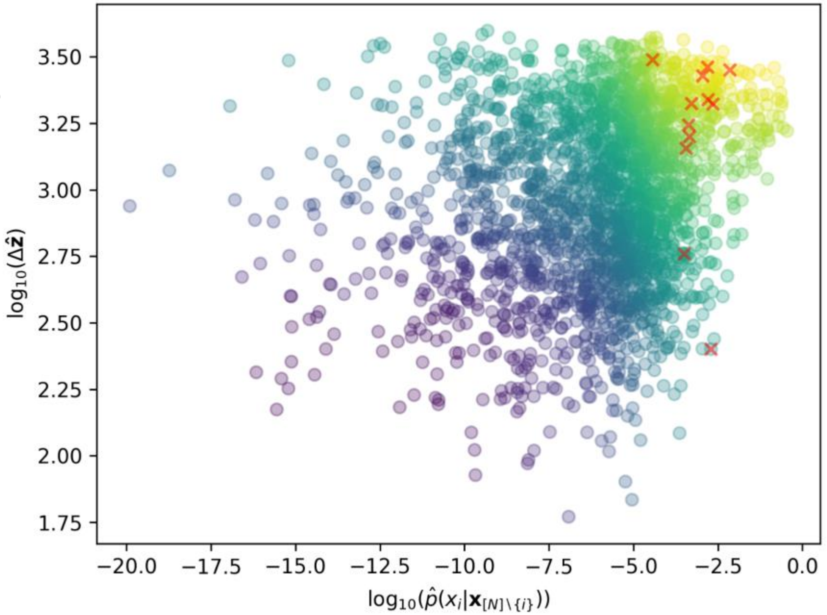
1. Learning the language of viral evolution and escape

2. Recovering evolutionary dynamics with "evolutionary velocity"

3. **Looking forward**

Predicting drug resistance

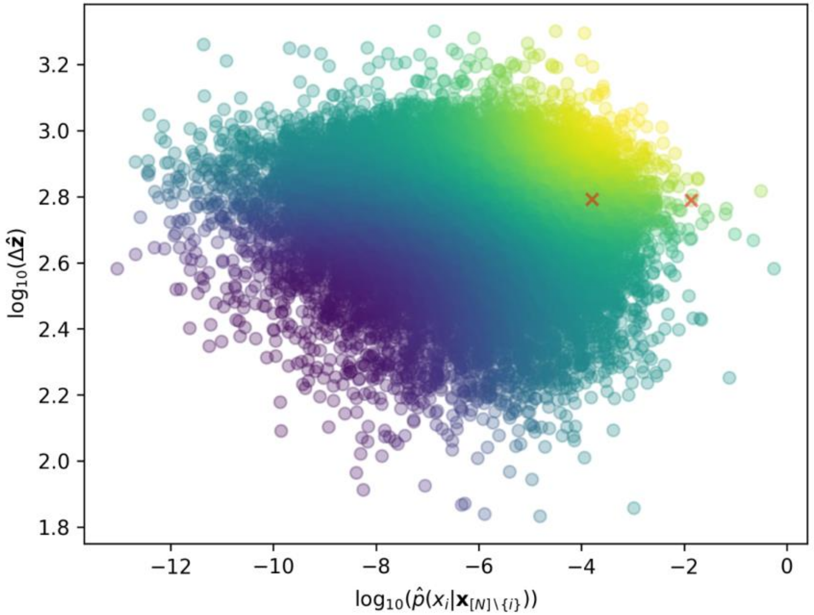
HIV



Semantic change

Grammaticality

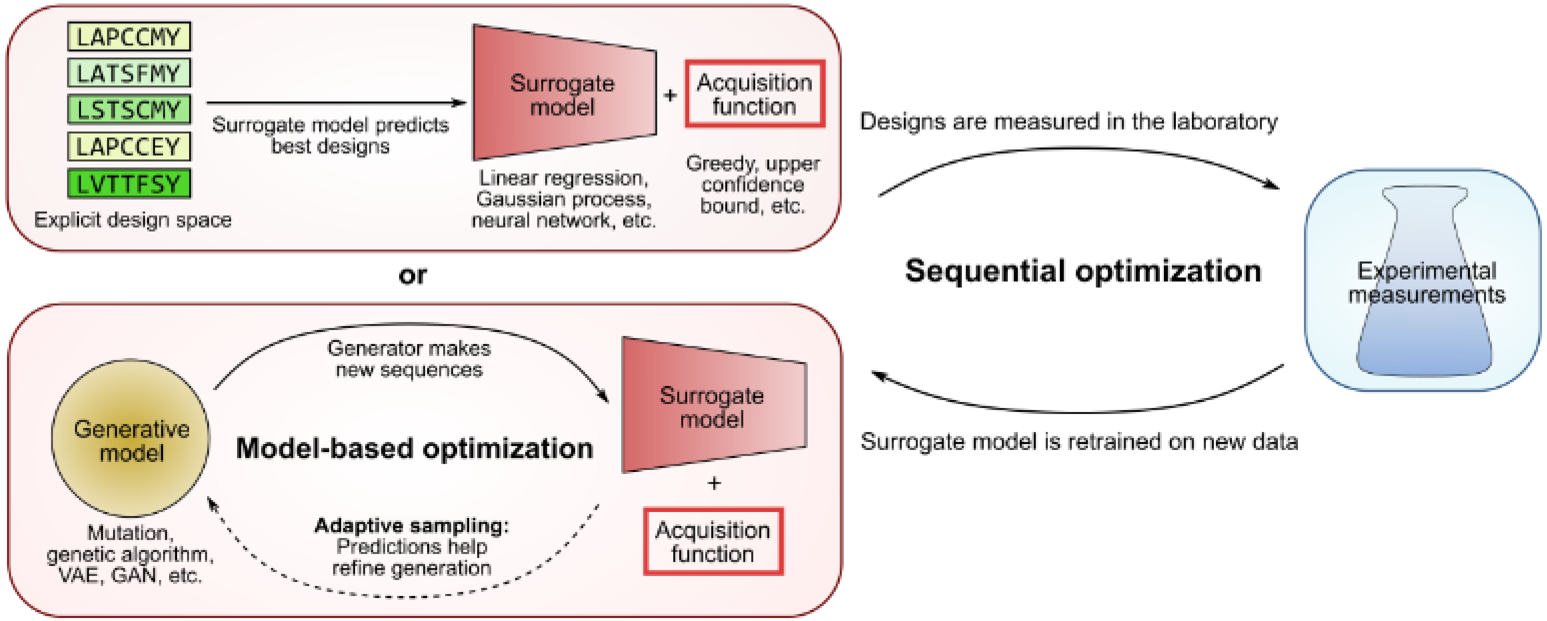
M. tuberculosis



Semantic change

Grammaticality

Evolving protein therapeutics


LAPCCMY
LATSFMY
LSTSCMY
LAPCCEY
LVTTFSY
Explicit design space
Surrogate model predicts best designs
Surrogate model
+
Acquisition function
Linear regression, Gaussian process, neural network, etc.
Greedy, upper confidence bound, etc.
or
Generative model
Generator makes new sequences
Model-based optimization
Surrogate model
+
Acquisition function
Mutation, genetic algorithm, VAE, GAN, etc.
Adaptive sampling: Predictions help refine generation
Designs are measured in the laboratory
Sequential optimization
Experimental measurements
Surrogate model is retrained on new data

From Hie and Yang; *arXiv*, 2021

Key takeaways

- Language models have the potential to improve evolutionary models and prediction

- Sufficient training data is important!

- Successful implementation will require interdisciplinary collaboration

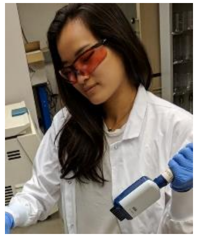
References

- Hie, Zhong, Berger, and Bryson.
  "Learning the language of viral evolution and escape."
  Science, 371:6526 (2021).

- Maher, …, Hie et al.
  "Predicting the mutational drivers of future SARS-CoV-2 variants of concern."
  medRxiv (2021).

- Hie, Yang, and Kim.
  "Evolutionary velocity with protein language models."
  bioRxiv (2021).

Thank you!



Ellen Zhong

MIT CSB CSAIL

Bryan Bryson

MIT MIT BE Ragon Institute of MGH, MIT and Harvard

Bonnie Berger

MIT MIT MATHEMATICS CSAIL

Kevin Yang

Microsoft Research

Peter Kim

Stanford MEDICINE | Biochemistry

Stanford ChEM-H

CHAN ZUCKERBERG INITIATIVE

Funding:

NDSEG

NIH National Institutes of Health

Stanford | Science Fellows