Protein Language Models

Sarah Gurev

MLCB 2024

Some slides adapted from Aaron Kollasch, Noor Youssef, John Ingraham, Adam Riesselman

Parallels between natural languages & proteins

DNA / RNA *4 nucleotides or 64 codons*	**Protein** *20 amino acids*
ATGTTCATCGTCCTG…	MADRLYMTKIHHQFDGD…

Parallels between natural languages & proteins

	Languages *26 alphabet*	**DNA / RNA** *4 nucleotides or 64 codons*	**Protein** *20 amino acids*
String	The cat sat on the mat.	ATGTTCATCGTCCTG…	MADRLYMTKIHHQFDGD…

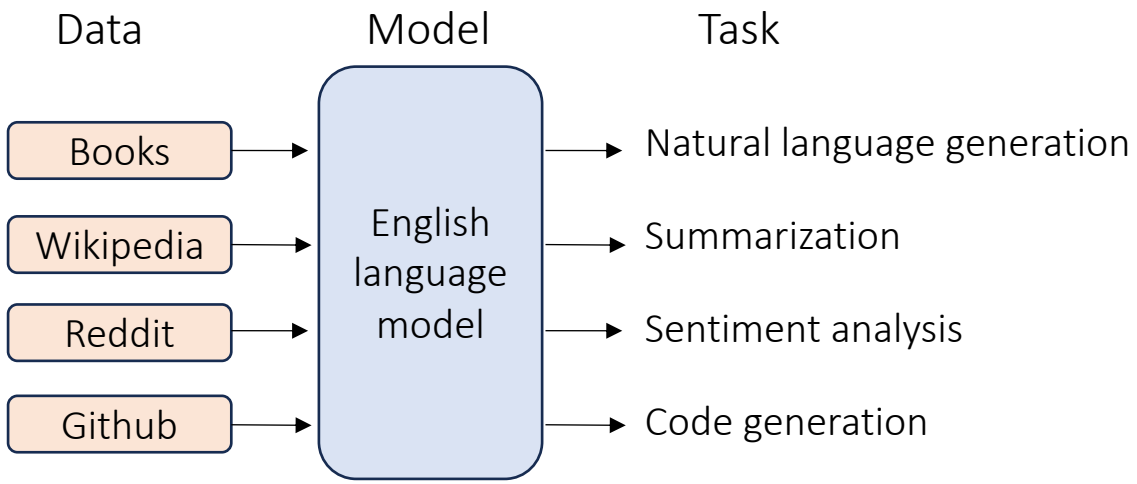
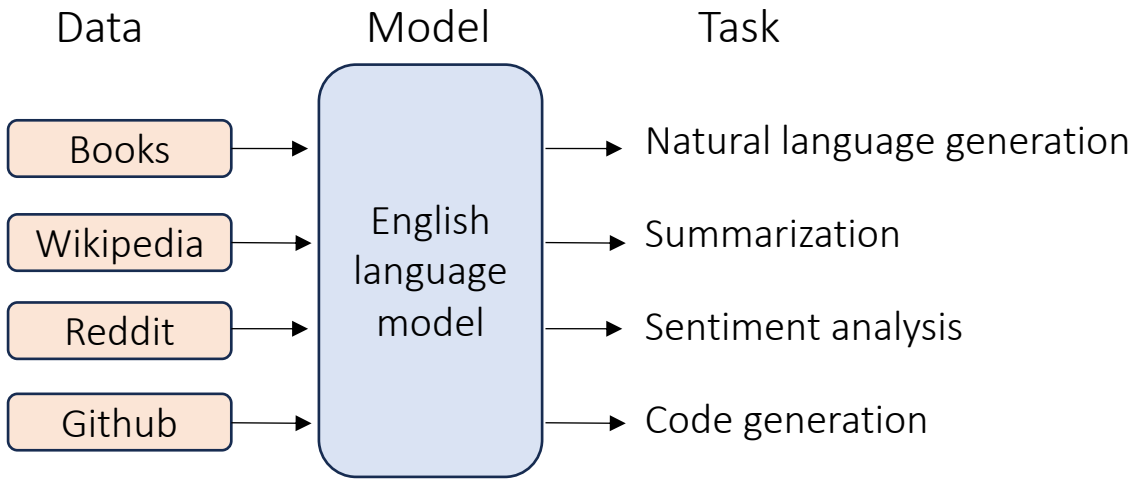


Figure adapted from Madani et al., 2023. Nature Biotech



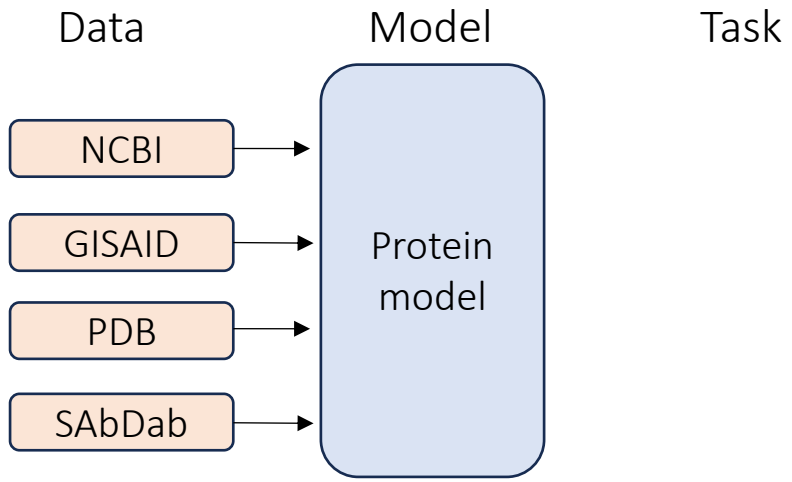


Figure adapted from Madani et al., 2023. Nature Biotech

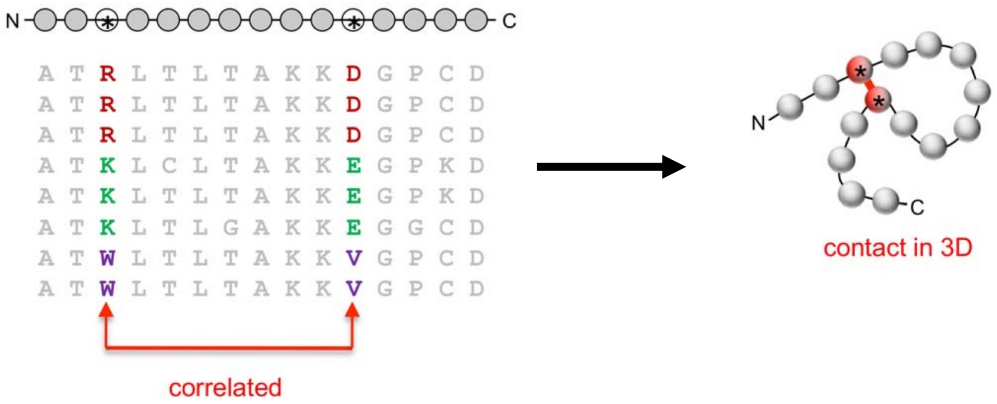
OPEN ACCESS Freely available online

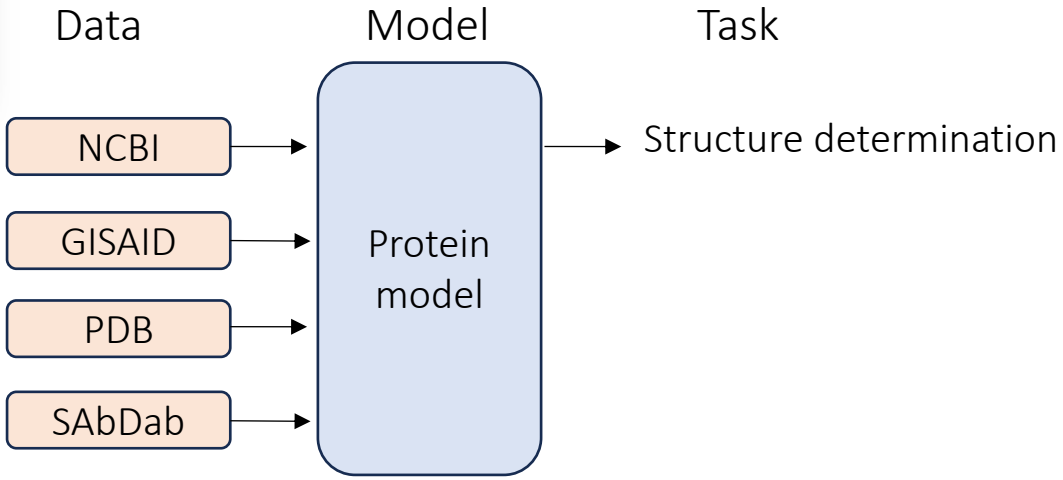
PLOS one

Protein 3D Structure Computed from Evolutionary Sequence Variation

Debora S. Marks[1*], Lucy J. Colwell[2], Robert Sheridan[3], Thomas A. Hopf[1], Andrea Pagnani[4], Riccardo Zecchina[4,5], Chris Sander[3]

1 Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America, **2** MRC Laboratory of Molecular Biology, Hills Road, Cambridge, United Kingdom, **3** Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, **4** Human Genetics Foundation, Torino, Italy, **5** Politecnico di Torino, Torino, Italy





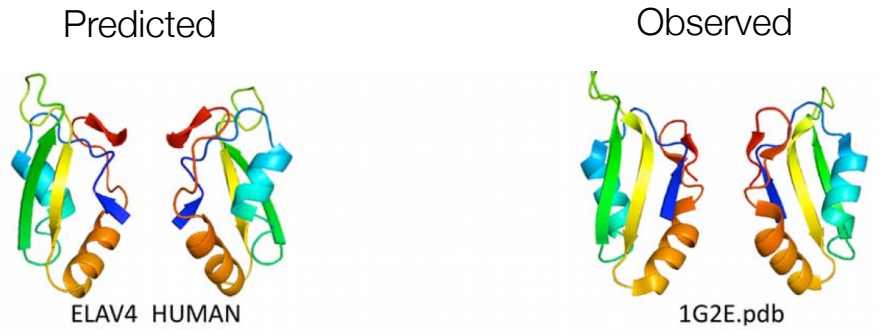
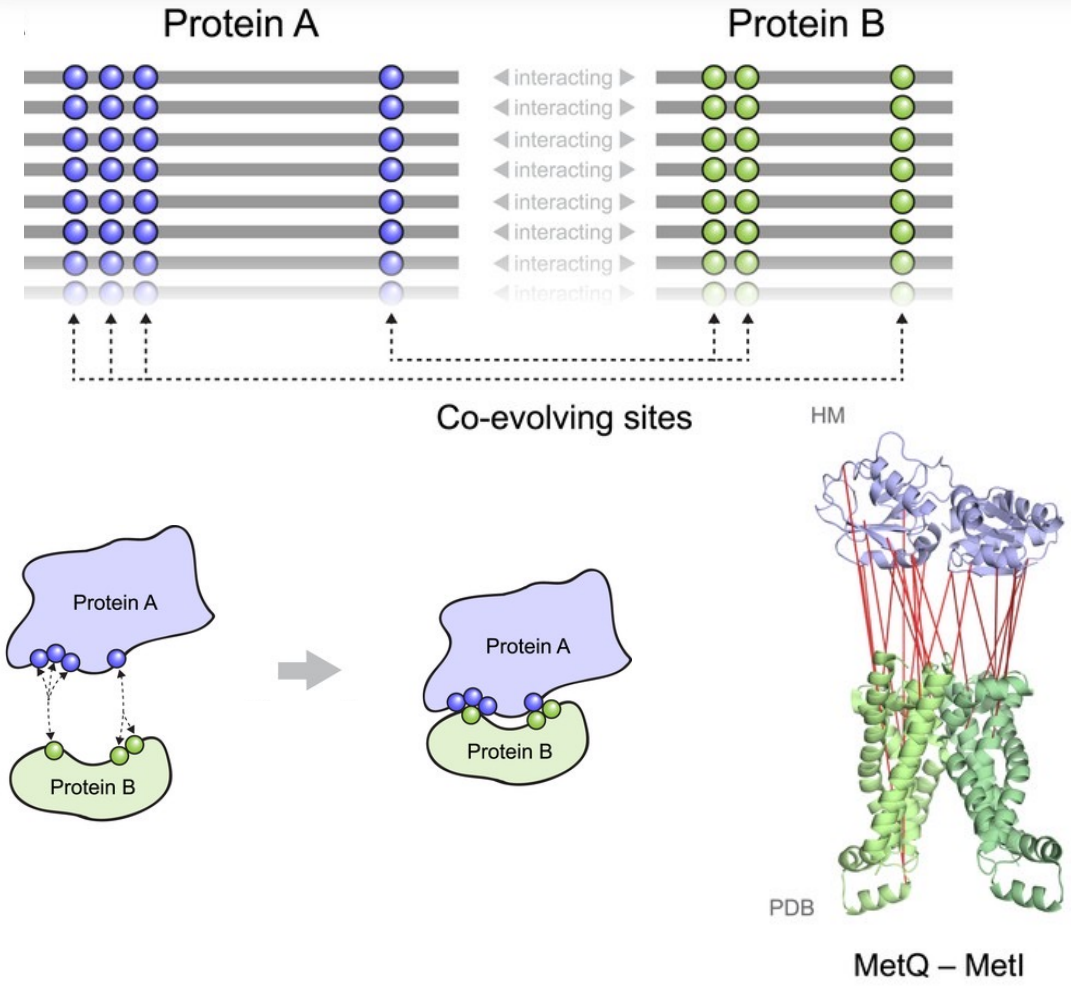


Figure adapted from Madani et al., 2023. Nature Biotech

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A Hopf[1,2†], Charlotta P I Schärfe[1,3,4†], João P G L M Rodrigues[5†], Anna G Green[1], Oliver Kohlbacher[3,4], Chris Sander[6*], Alexandre M J J Bonvin[5*], Debora S Marks[1*]



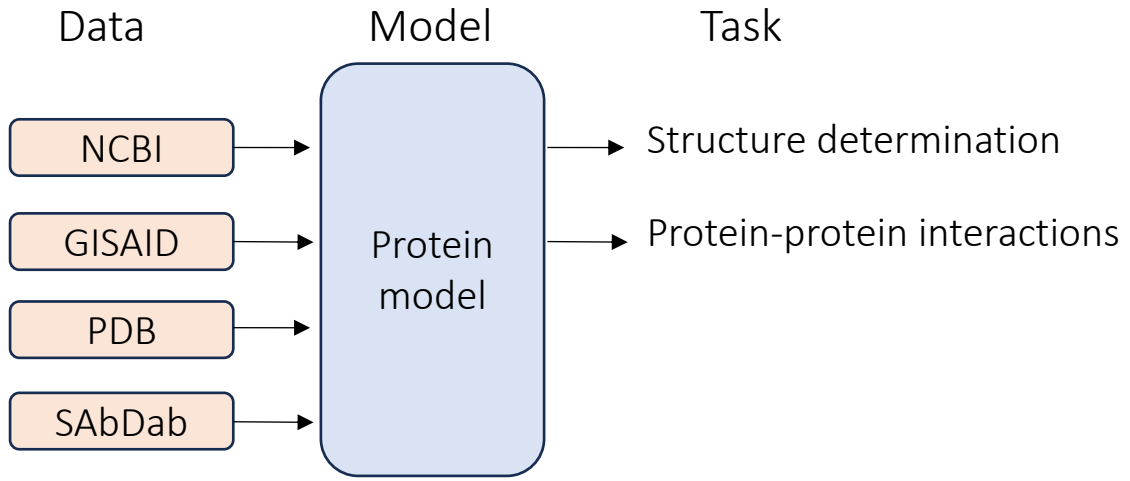


Figure adapted from Madani et al., 2023. Nature Biotech

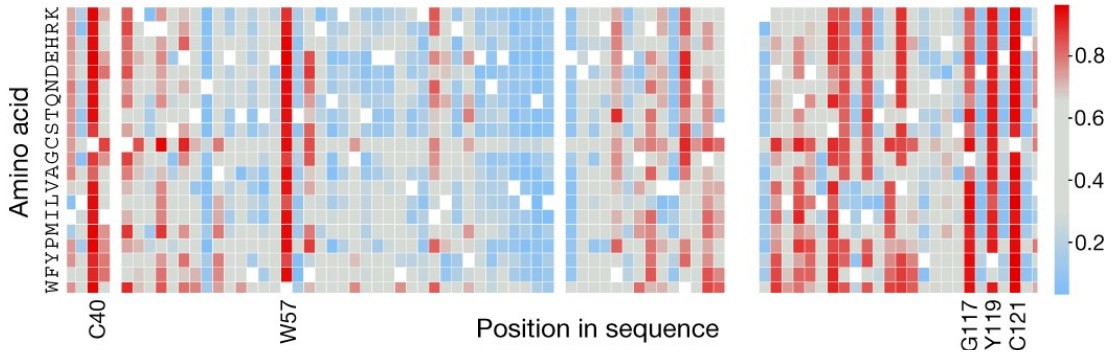
Article

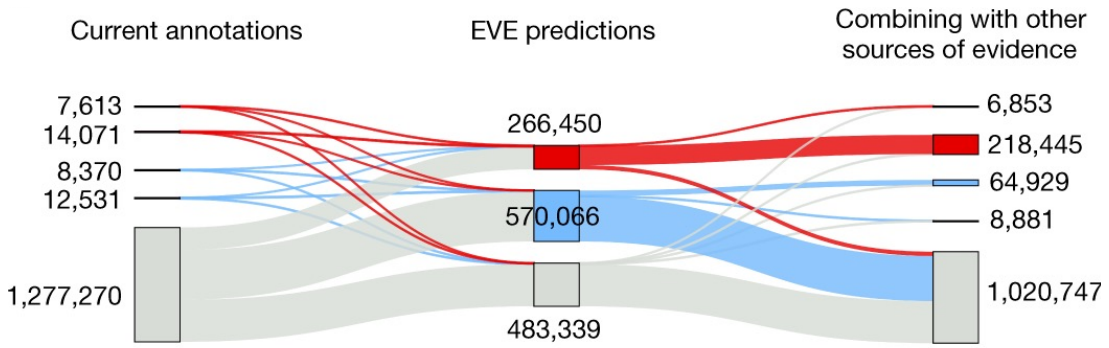
Disease variant prediction with deep generative models of evolutionary data

https://doi.org/10.1038/s41586-021-04043-8

Received: 18 December 2020

Jonathan Frazer[1,4], Pascal Notin[2,4], Mafalda Dias[1,4], Aidan Gomez[2], Joseph K. Min[1], Kelly Brock[1], Yarin Gal[2]✉ & Debora S. Marks[1,3]✉





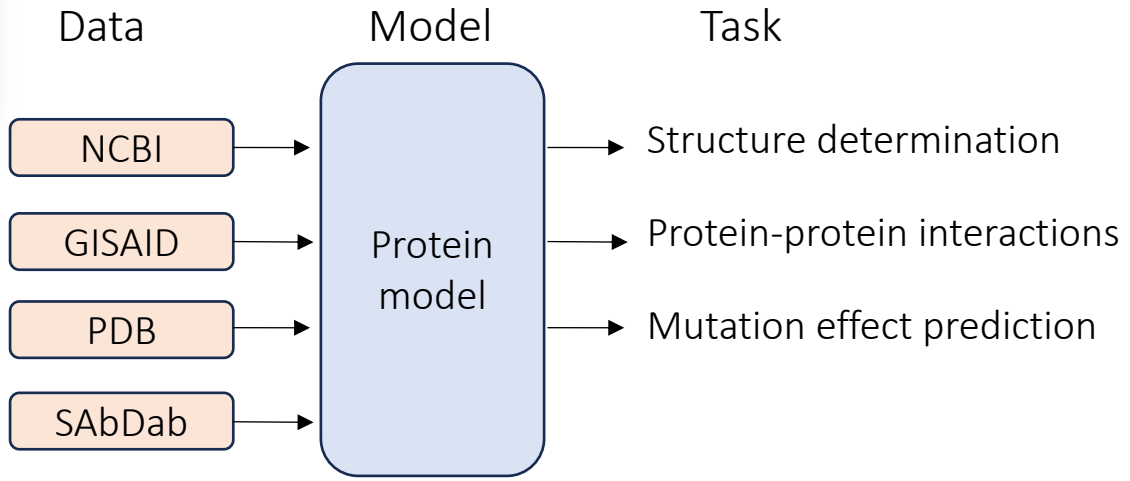


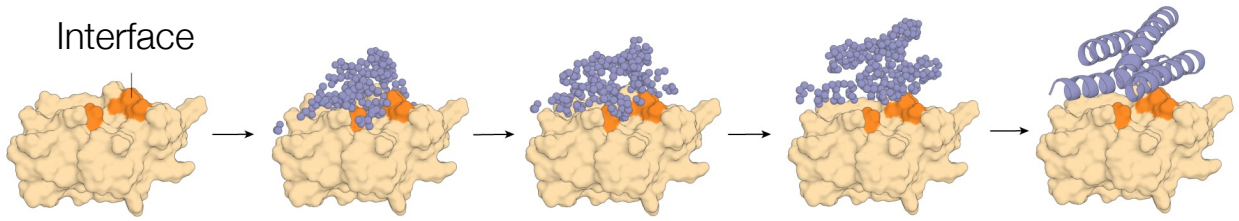
Figure adapted from Madani et al., 2023. Nature Biotech

Article

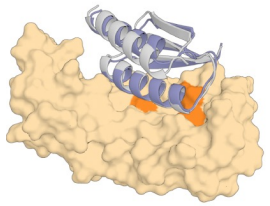
De novo design of protein structure and function with RFdiffusion

https://doi.org/10.1038/s41586-023-06415-8
Received: 14 December 2022
Accepted: 7 July 2023
Published online: 11 July 2023
Open access
Check for updates

Joseph L. Watson[1,2,15], David Juergens[1,2,3,15], Nathaniel R. Bennett[1,2,3,15], Brian L. Trippe[2,4,5,15], Jason Yim[2,6,15], Helen E. Eisenach[1,2,15], Woody Ahern[1,2,7,15], Andrew J. Borst[1,2], Robert J. Ragotte[1,2], Lukas F. Milles[1,2], Basile I. M. Wicky[1,2], Nikita Hanikel[1,2], Samuel J. Pellock[1,2], Alexis Courbet[1,2,8], William Sheffler[1,2], Jue Wang[1,2], Preetham Venkatesh[1,2,9], Isaac Sappington[1,2,9], Susana Vázquez Torres[1,2,9], Anna Lauko[1,2,9], Valentin De Bortoli[8], Emile Mathieu[10], Sergey Ovchinnikov[11,12], Regina Barzilay[6], Tommi S. Jaakkola[6], Frank DiMaio[1,2], Minkyung Baek[13] & David Baker[1,2,14]✉



Novel binder to influenza HA protein



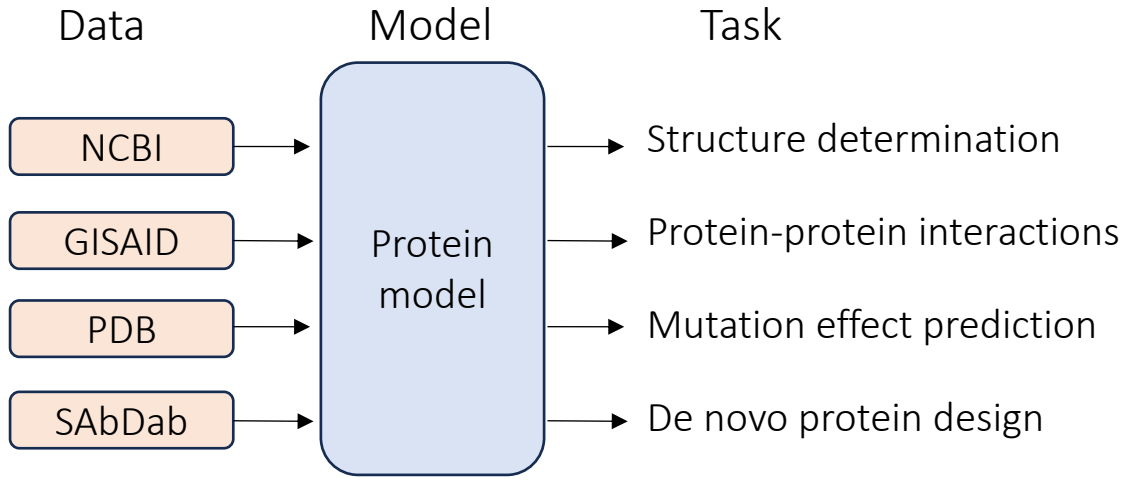
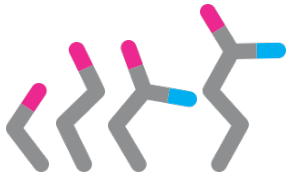


Figure adapted from Madani et al., 2023. Nature Biotech

The protein alphabet

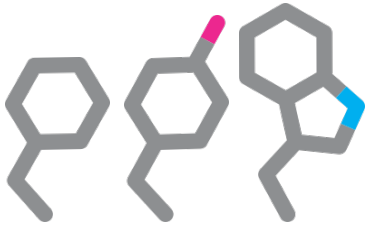
positive

negative



polar

“greasy”

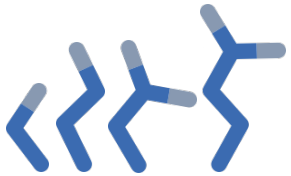


aromatic

other stuff

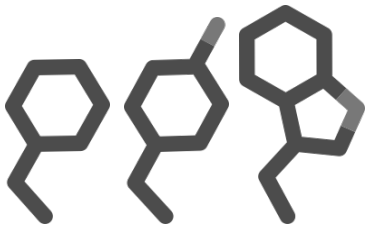
positive

negative



polar

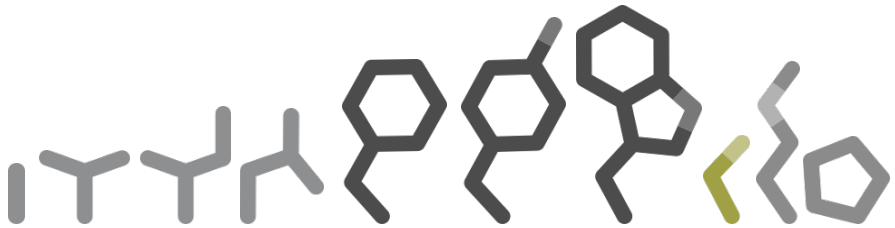
“greasy”



aromatic

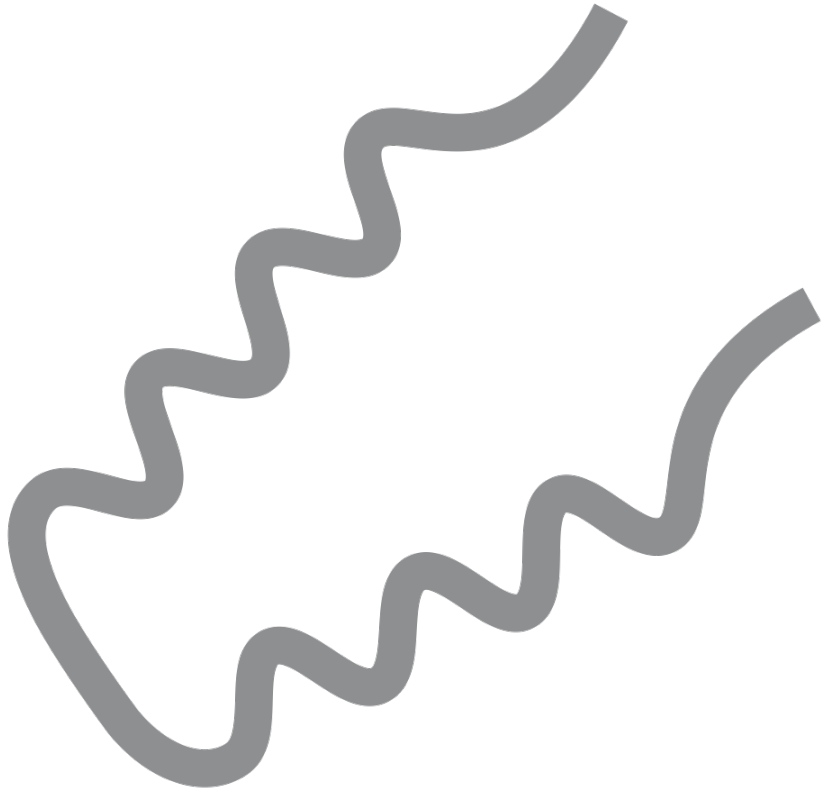
other stuff

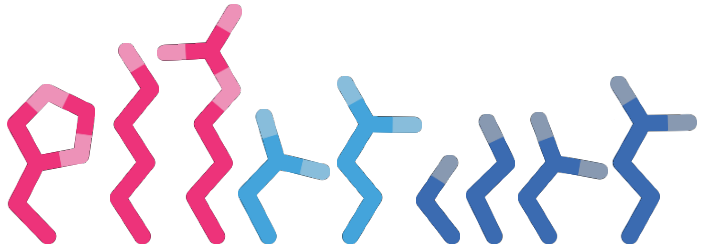
hydrophilic



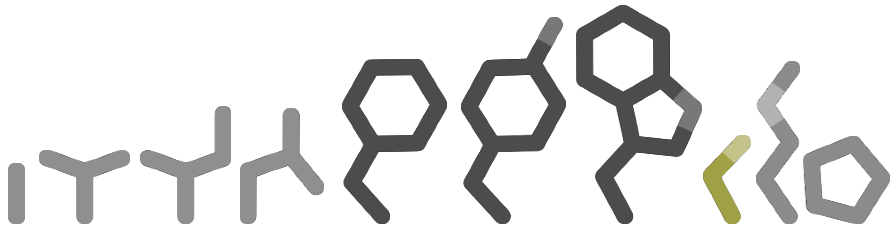
hydrophobic

Slide from Ingrahm & Riesselman, MIA 2017



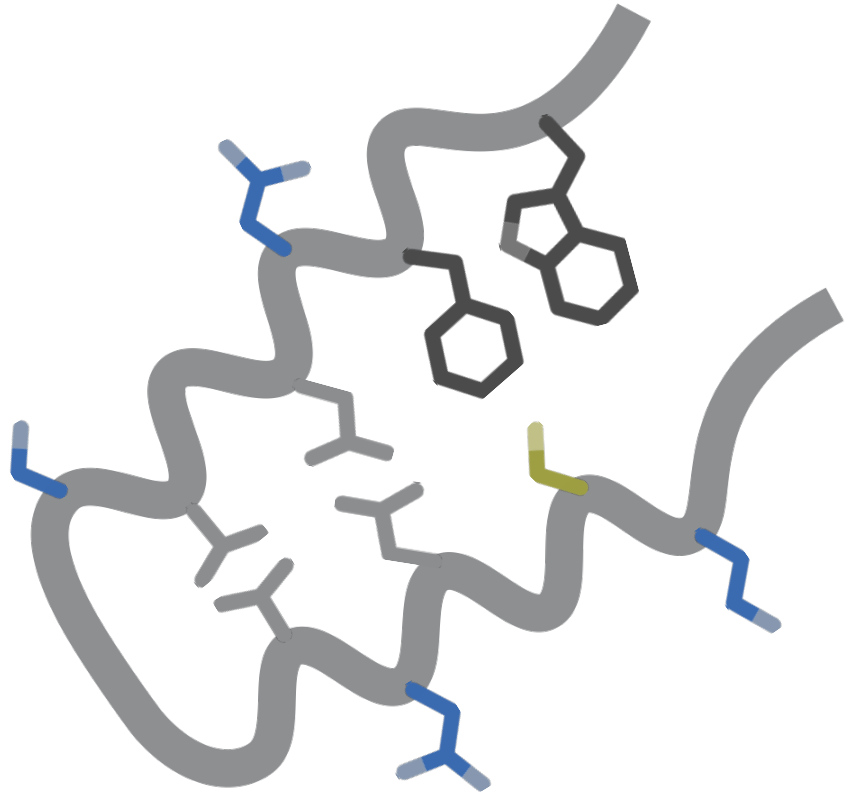


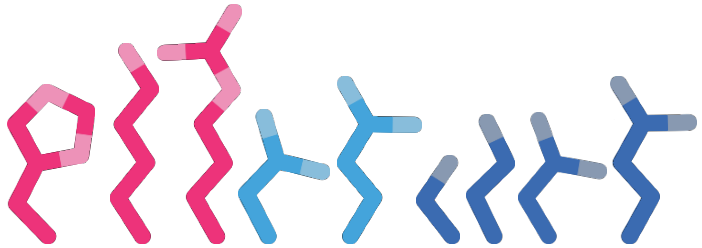
hydrophilic



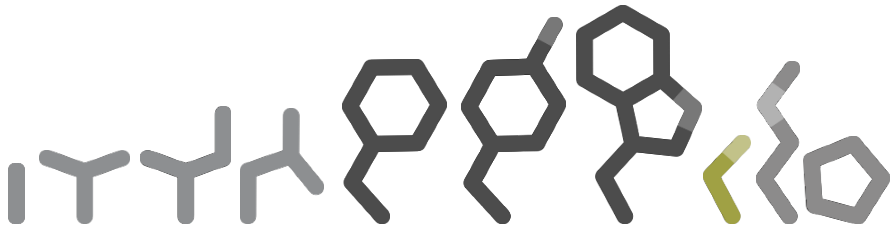
hydrophobic

Slide from Ingrahm & Riesselman, MIA 2017



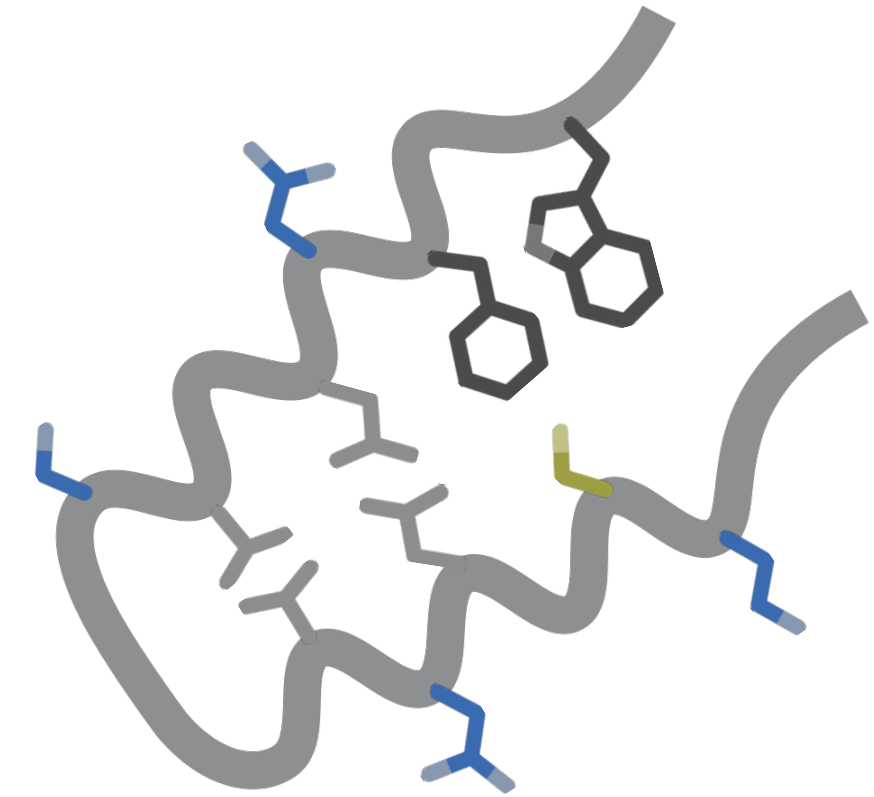


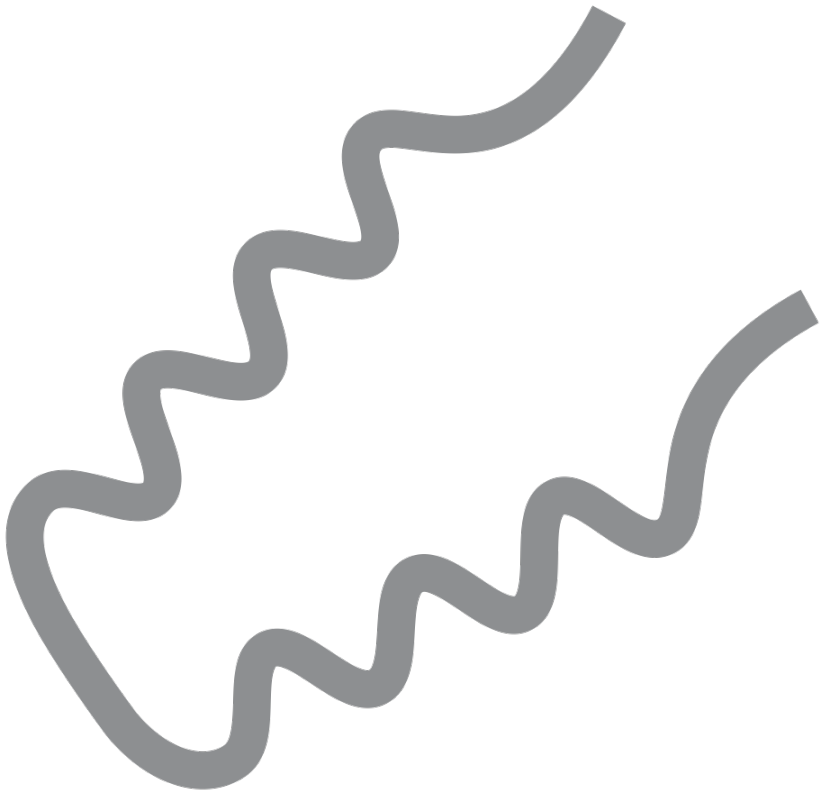
hydrophilic

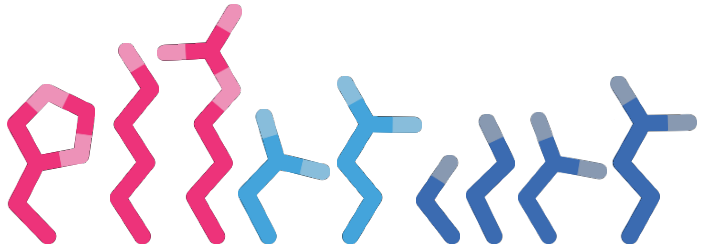


hydrophobic

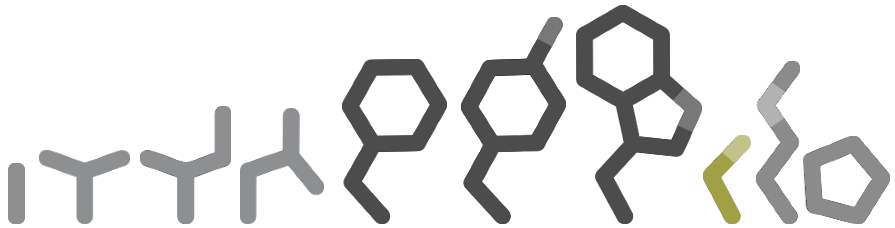
Slide from Ingrahm & Riesselman, MIA 2017





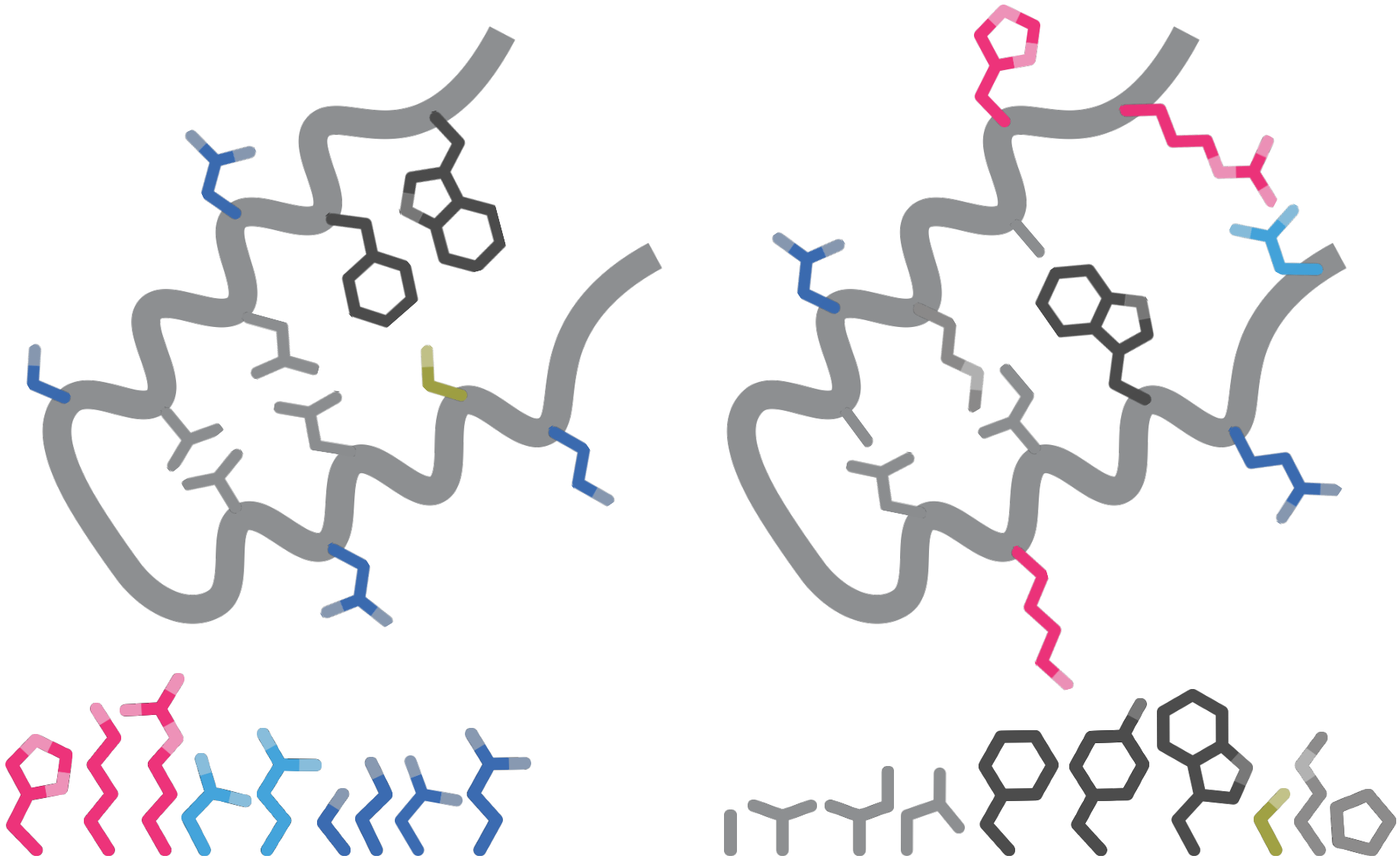


hydrophilic

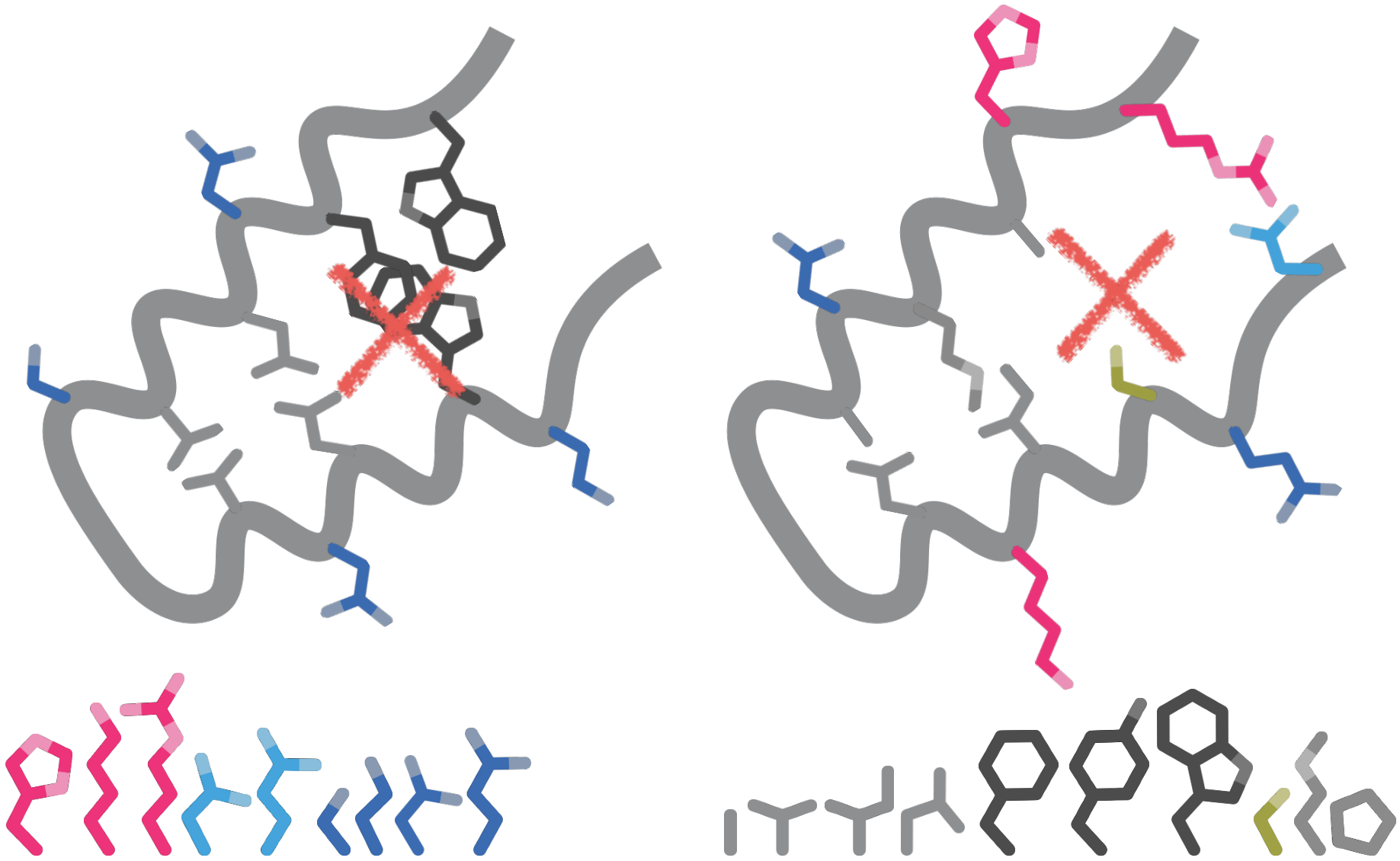


hydrophobic

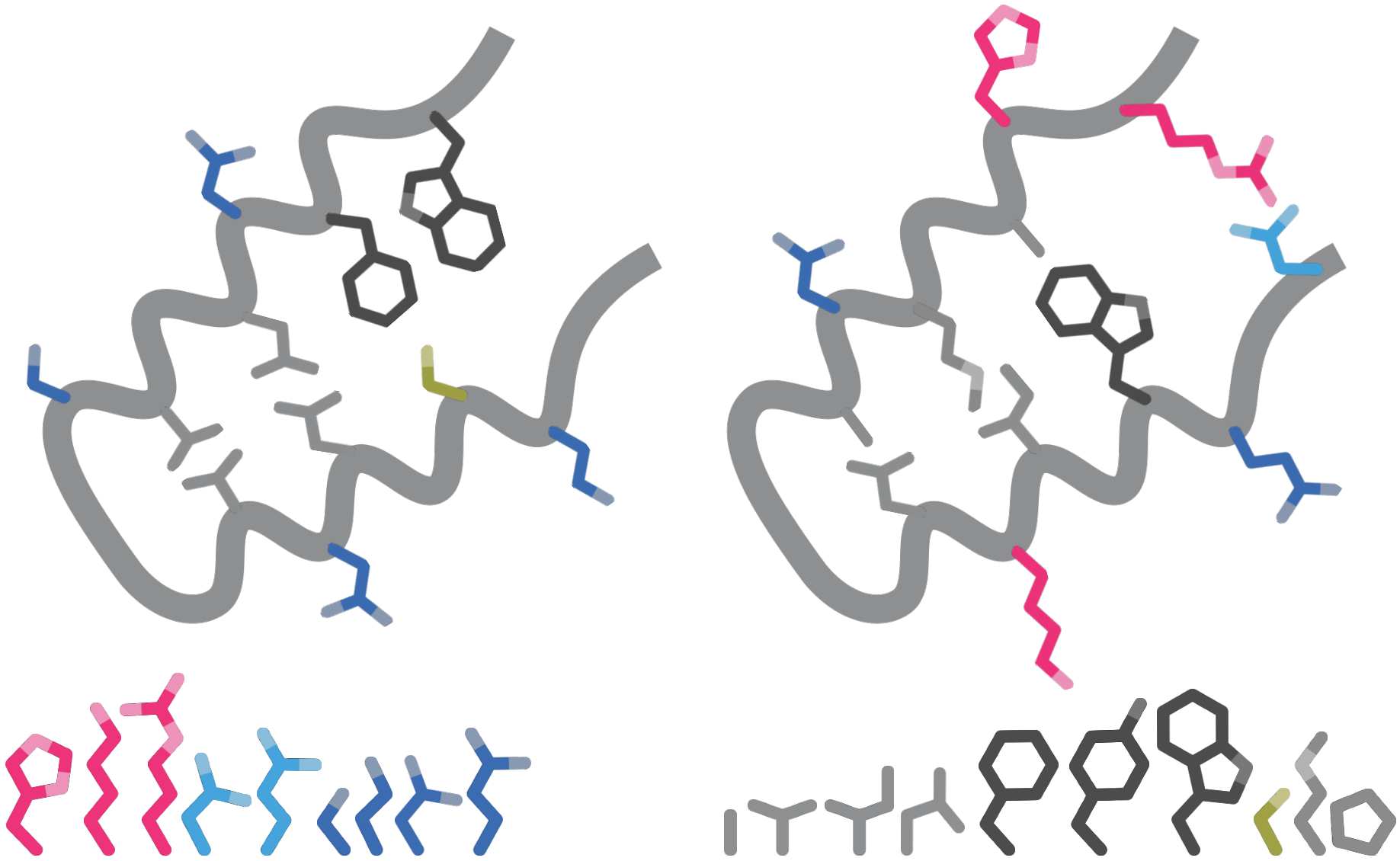
Slide from Ingrahm & Riesselman, MIA 2017



Slide from Ingrahm & Riesselman, MIA 2017

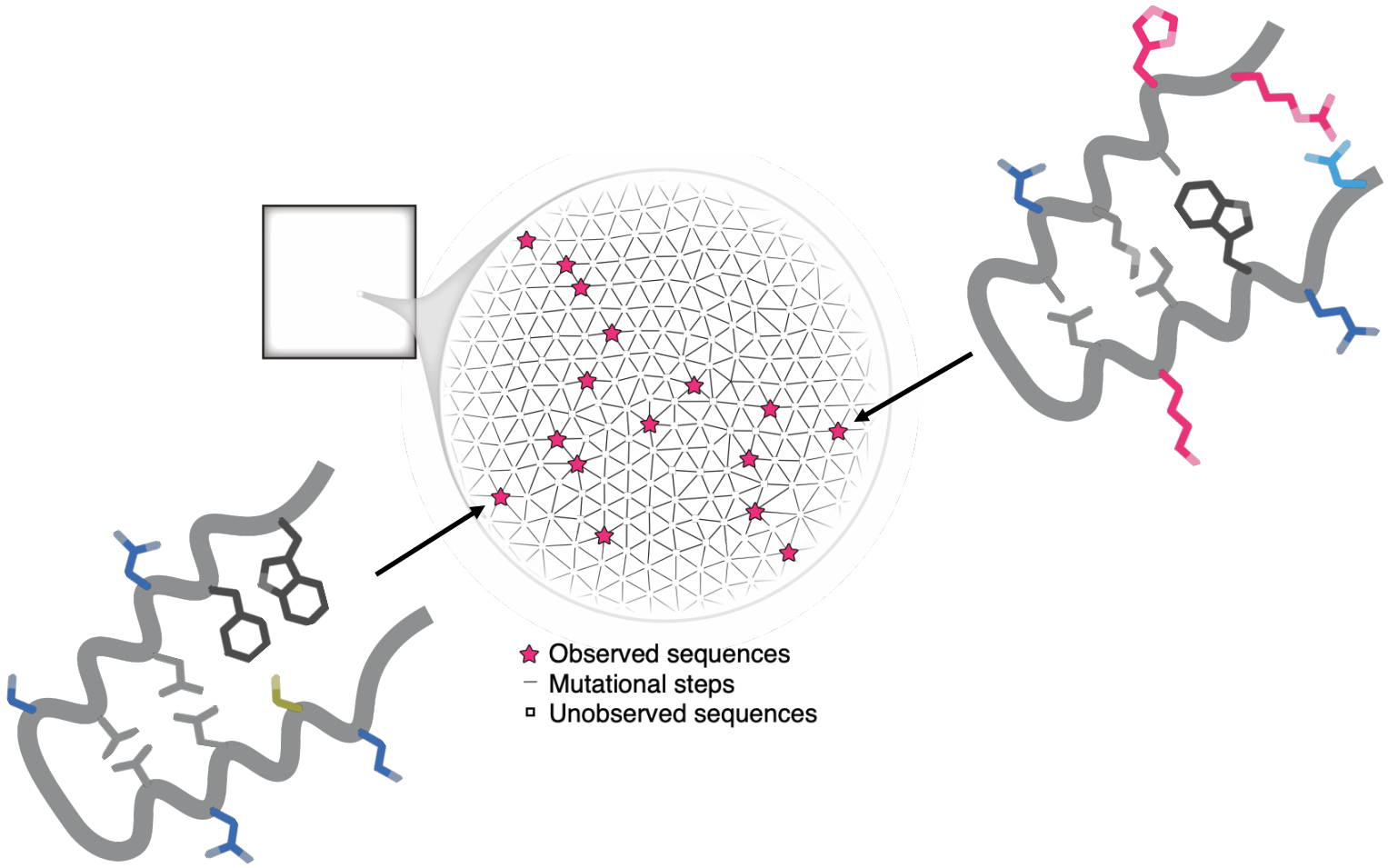


Slide from Ingrahm & Riesselman, MIA 2017



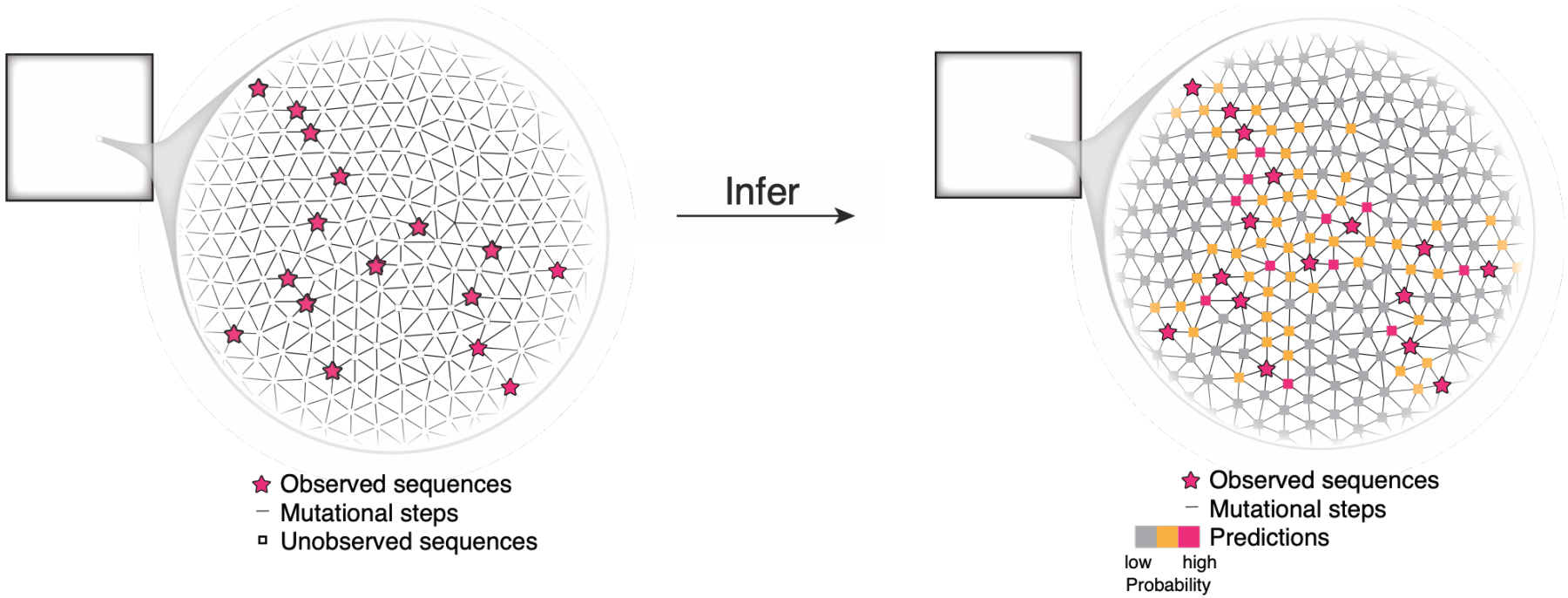
Slide from Ingrahm & Riesselman, MIA 2017

Evolution samples functional sequences



Slide from Ingrahm & Riesselman, MIA 2017

We want a model that can learn the "patterns" of functional sequences



Slide from Ingrahm & Riesselman, MIA 2017

The data: multiple *sequence* alignments

- AQKLYLTHIDAEVEGD
- ADRLYMTKIHHQFDGD
- ADTLFITEVKQVFEGD
- ADRLYMTKIHHTFDGD
- ADKLYCTLIHNSFEGD
- ADRLYMTKIHHEFEGD
- ADRLYLTMIHQKFEAD
- TDRLYITHIDETFEGD
- ADRLYLTQIRNKFKGD
- TSKMYITKIGQEFEGD
- ADRLYMTKIHHEFEGD
- ADRLYITHIHHSFEGD
- ADRLYMTKIHHEFEGD

- Structure and functional constraints is encoded in sequences
- Evolutionarily-related sequences can be aligned

Slide from Ingrahm & Riesselman, MIA 2017

Many ways to model sequences

- Site-independent models
- Potts model (pairwise interactions)
- Higher order interactions

Many ways to model sequences

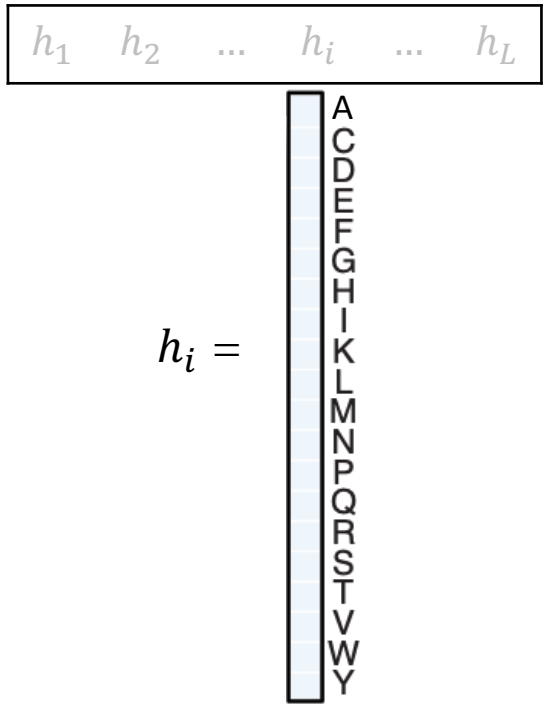
- Site-independent models
- Potts model (pairwise interactions)
- Higher order interactions

Site independence model

Multiple sequence alignment

Position i

Position specific amino acid frequencies



Score sequences

$\boldsymbol{x}$ = ADRLYMTKIHHQFDGT

Product of **site** factors

$$P(\boldsymbol{x}) = p_1(\mathrm{A})p_2(\mathrm{D})p_3(\mathrm{R})\ldots p_L(\mathrm{T})$$

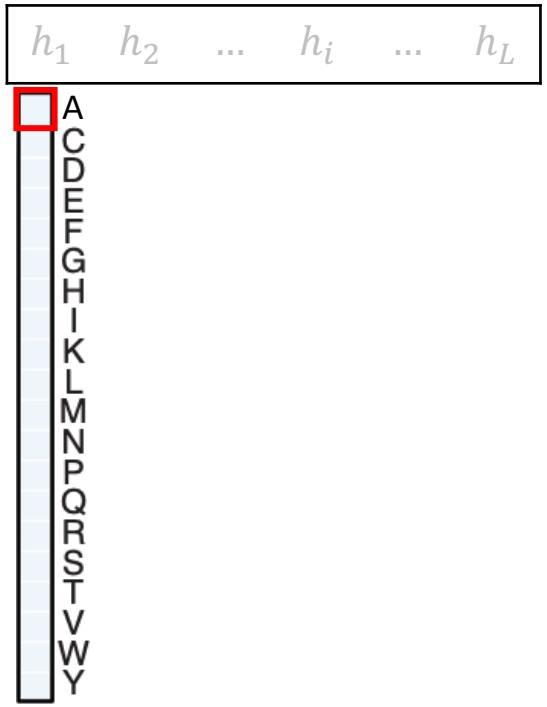
Slide from Ingrahm & Riesselman, MIA 2017

Site independence model

Multiple sequence alignment

Position i

Position specific amino acid frequencies



Score sequences

$\boldsymbol{x}$ = ADRLYMTKIHHQFDGT

Product of **site** factors

$$P(\boldsymbol{x}) = p_1(\mathrm{A})p_2(\mathrm{D})p_3(\mathrm{R})\ldots p_L(\mathrm{T})$$

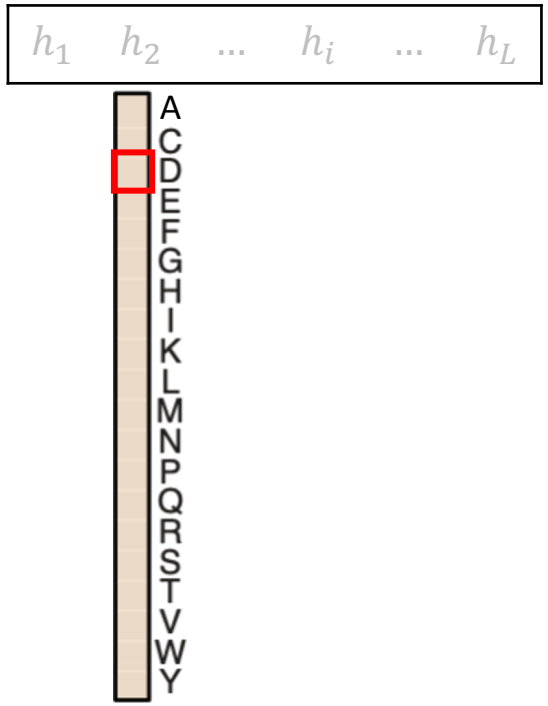
Slide from Ingrahm & Riesselman, MIA 2017

Site independence model

Multiple sequence alignment

Position i

Position specific amino acid frequencies



Score sequences

$\boldsymbol{x}$ = ADRLYMTKIHHQFDGT

Product of **site** factors

$$P(\boldsymbol{x}) = p_1(\mathrm{A})p_2(\mathrm{D})p_3(\mathrm{R})\ldots p_L(\mathrm{T})$$

Slide from Ingrahm & Riesselman, MIA 2017

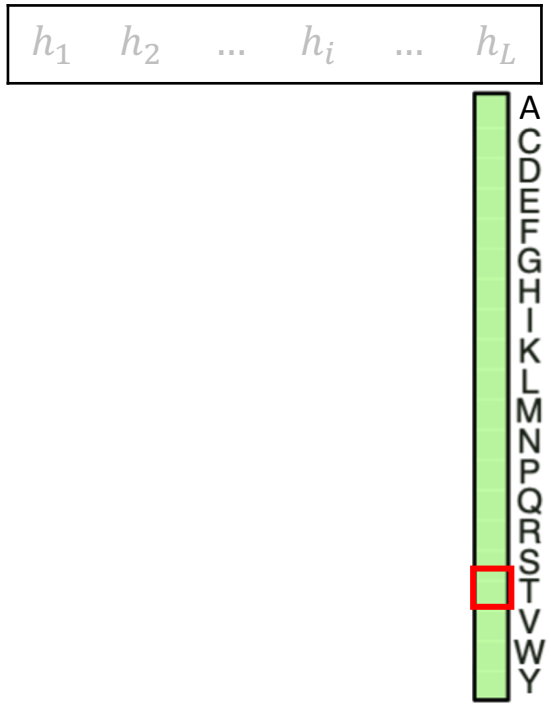
Site independence model

Multiple sequence alignment

```
AQKLYLTHIDAEVEGD
ADRLYMTKIHHQFDGD
ADTLFITEVKQVFEGD
ADRLYMTKIHHTFDGD
ADKLYCTLIHNSFEGD
ADRLYMTKIHHEFEGD
ADRLYLTMIHQKFEAD
TDRLYITHIDETFEGD
ADRLYLTQIRNKFKGD
TSKMYITKIGQEFEGD
ADRLYMTKIHHEFEGD
ADRLYITHIHHSFEGD
ADRLYMTKIHHEFEGD
```

Position i

Position specific amino acid frequencies



Score sequences

$\boldsymbol{x}$ = ADRLYMTKIHHQFDGT

Product of **site** factors

$$P(\boldsymbol{x}) = p_1(\mathrm{A})p_2(\mathrm{D})p_3(\mathrm{R}) \ldots p_L(\mathrm{T})$$

$$= p_1(x_1)p_2(x_2) \ldots p_L(x_L)$$

or equivalently

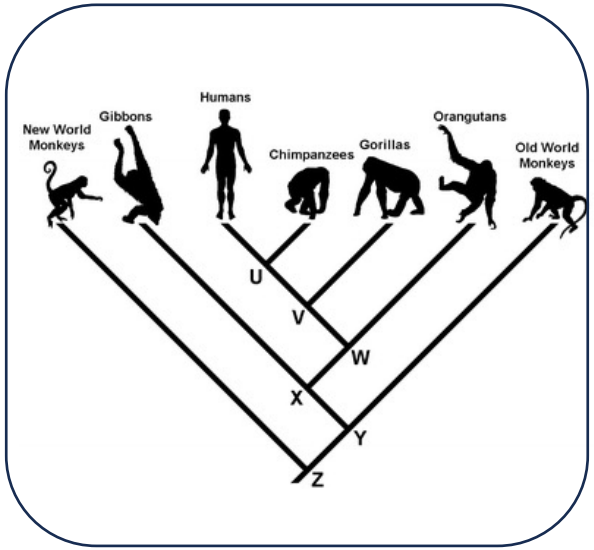
$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_i h_i(x_i)\right)$$

$$h_i = \boldsymbol{log}(p_i(x_i))$$

Slide from Ingrahm & Riesselman, MIA 2017

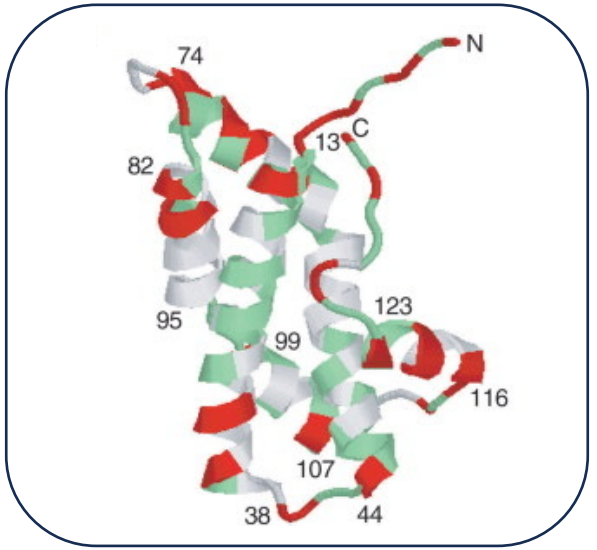
Site independence model

Phylogenetic inference



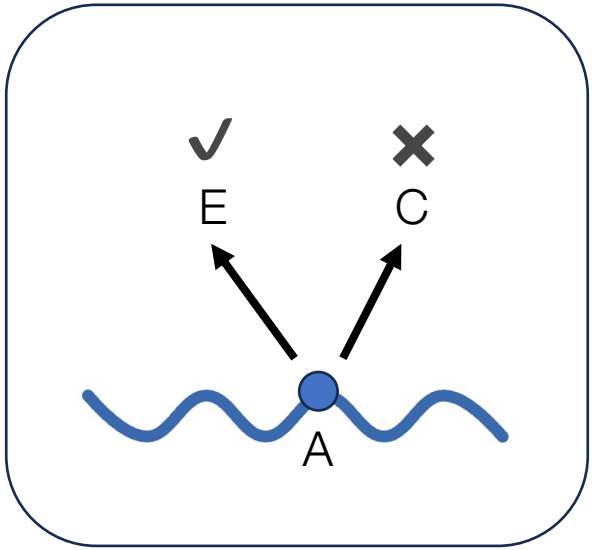
Ronquist *et al* 2012 *Syst Biol*
Suchard *et al* 2018 *Virus Evol*
Minh *et al* 2020 *Mol Biol Evol*

Substitution rate estimates



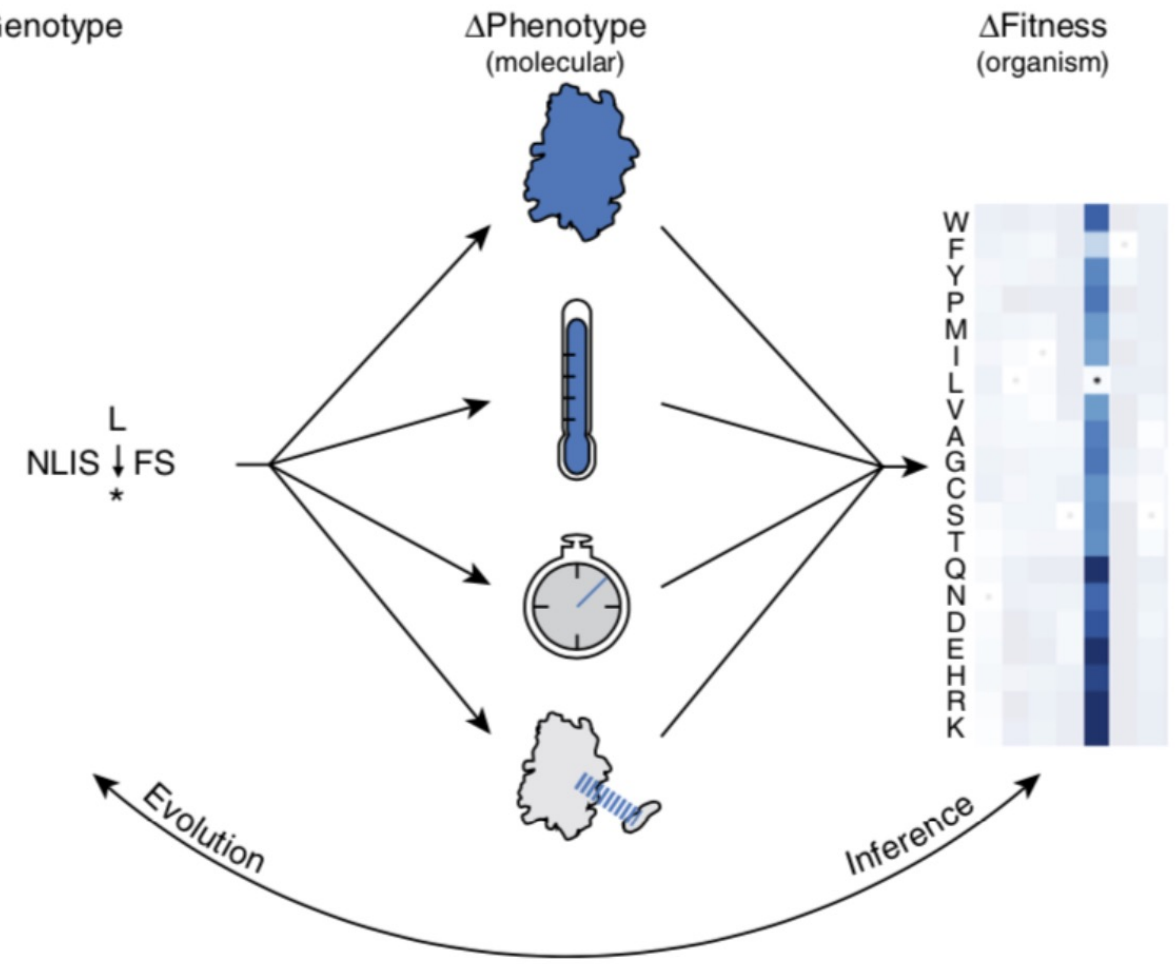
Pond *et al* 2020 *Mol Biol Evol*
Jones, Youssef *et al* 2020 *Syst Biol*
Youssef *et al* 2020 *Mol Biol Evol*

Mutation effect prediction



Hopf *et al* 2017 *Nature Biotech*
Riesselman *et al* 2018 *Nature Meth*

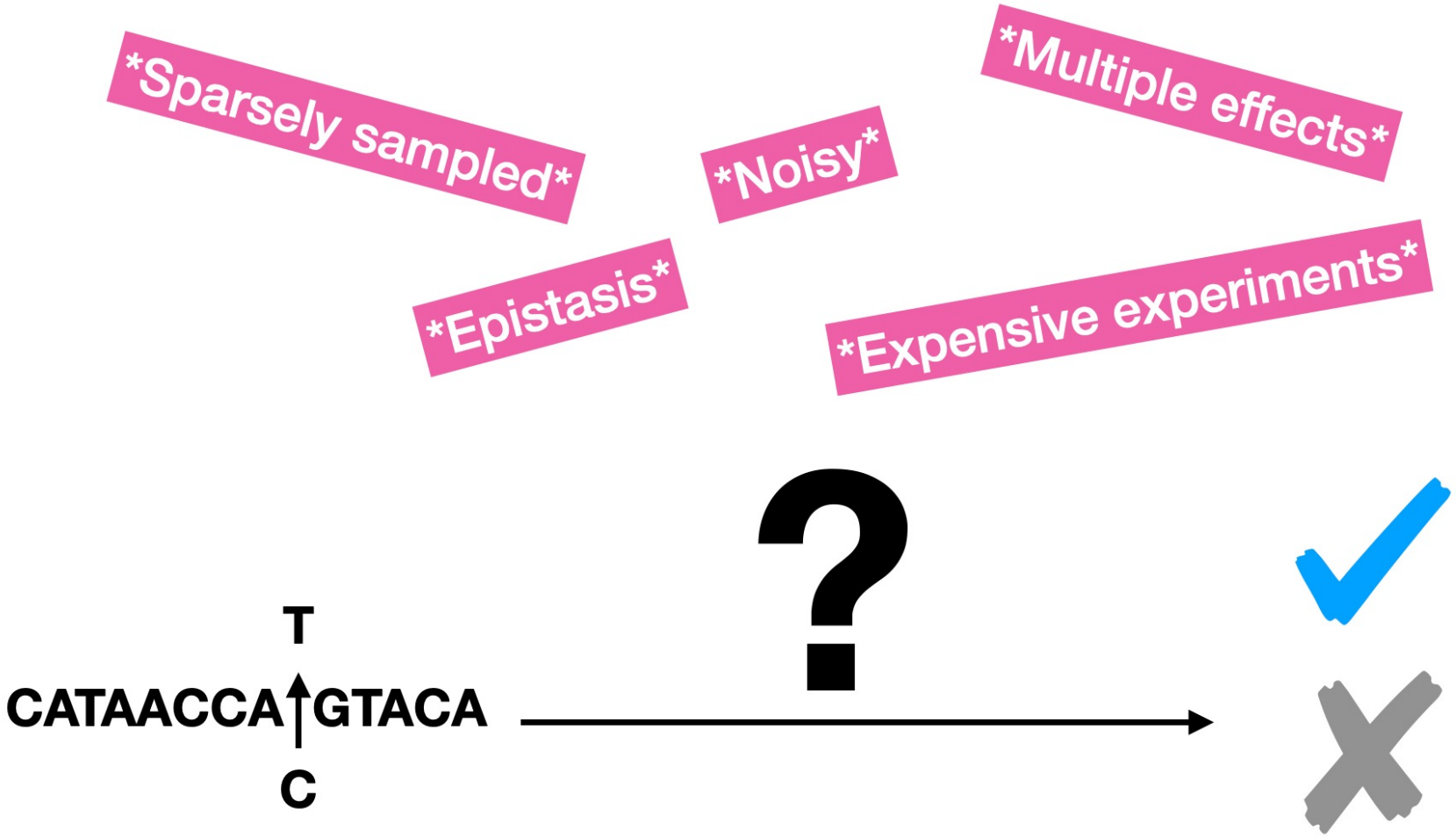
Mutation effect prediction: Mutations alter biological function



Slide from Ingraham & Riesselman, MIA 2017

Slide from Aaron Kollasch

Mutation effect prediction is hard!



Slide from Aaron Kollasch

Site independence model

Most common name

Mohammed	**James**	**Wang**
First name	Middle name	Last name

Epistasis: Sequence context matters!

Many ways to model sequences

- Site-independent models
- Potts model (pairwise interactions)
- Higher order interactions

Pairwise interactions model

Multiple sequence alignment

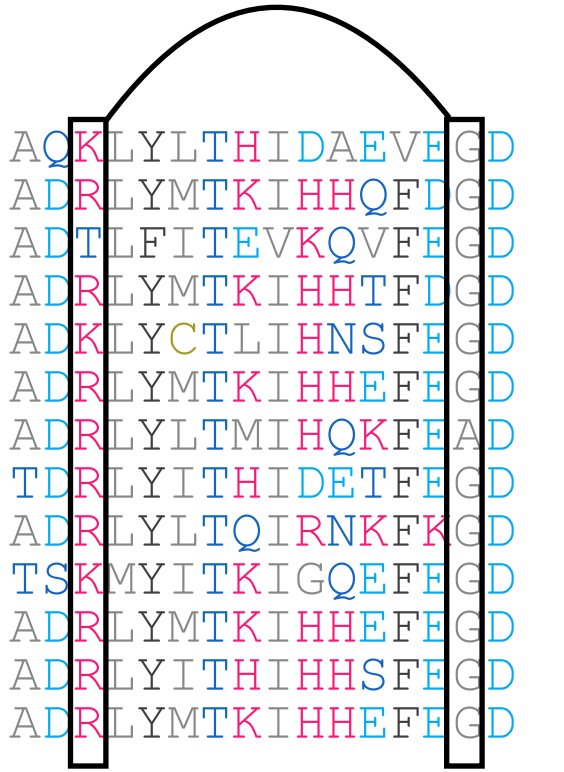
Position i Position j

Position specific amino acid frequencies

h_1 h_2 ... h_i ... h_L

Pairwise interactions model

Multiple sequence alignment

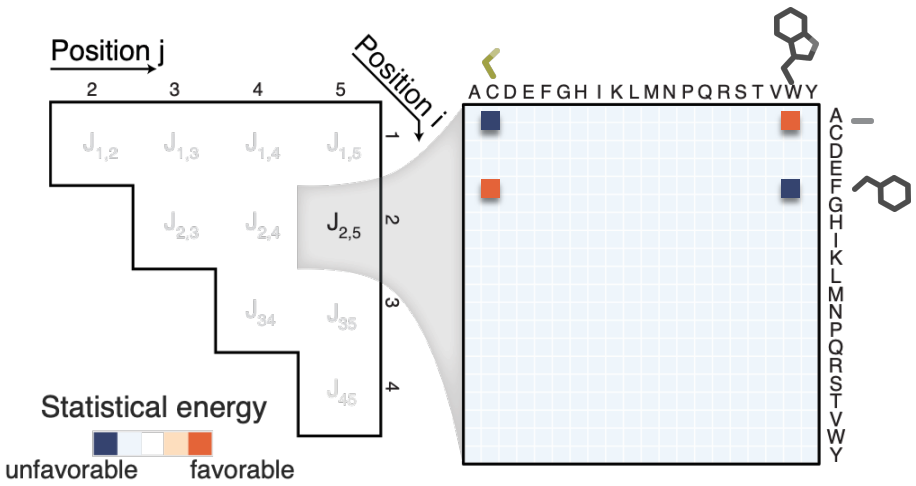


Position i Position j

Position specific amino acid frequencies

$$h_1 \quad h_2 \quad \ldots \quad h_i \quad \ldots \quad h_L$$

Pairwise frequencies



Score sequences

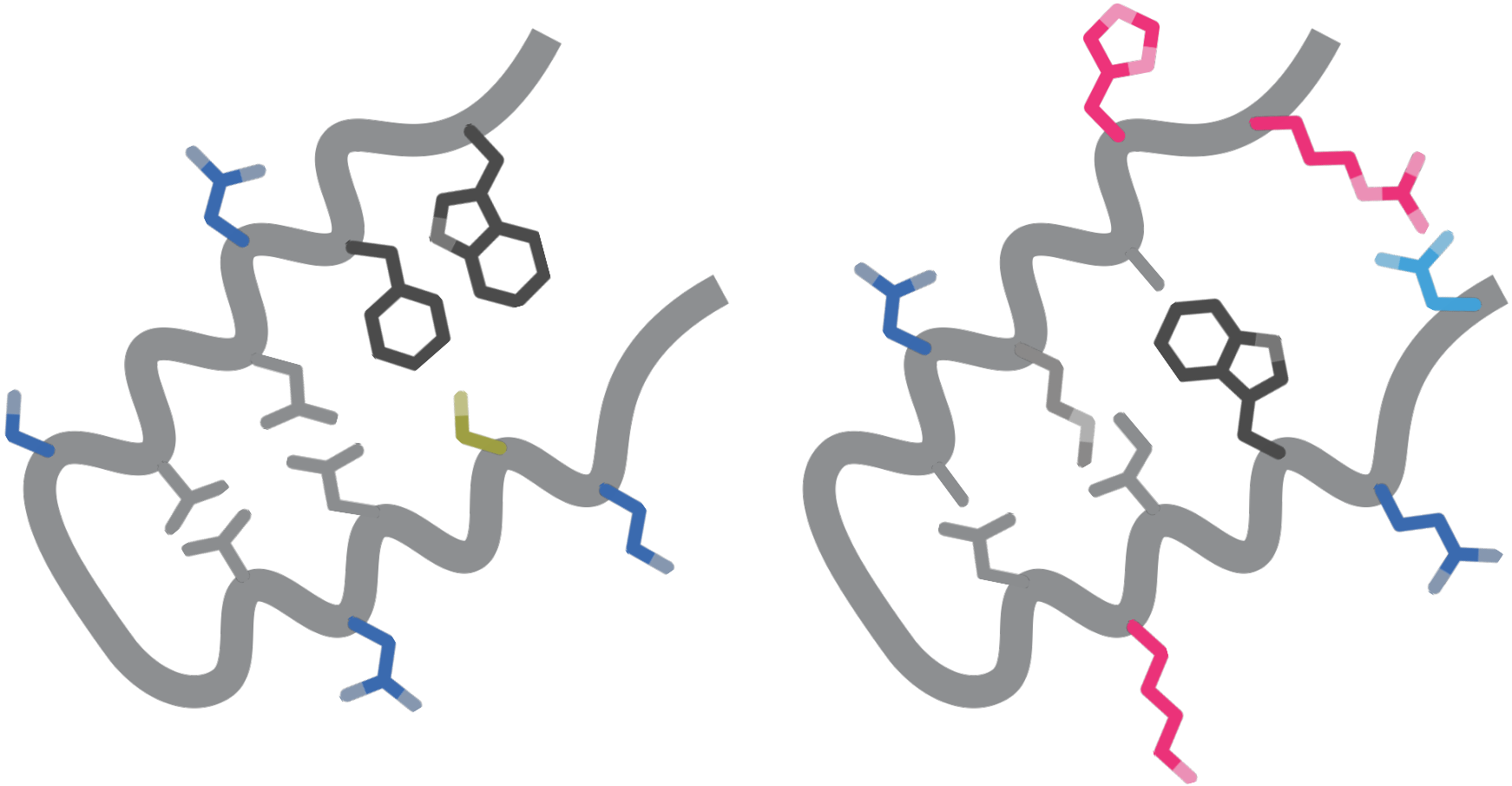
ADRLYMTKIHHQFDGD

Product of **site** and **pair** factors

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_i h_i(x_i) + \sum_{i<j} j_{ij}(x_i, x_j)\right)$$

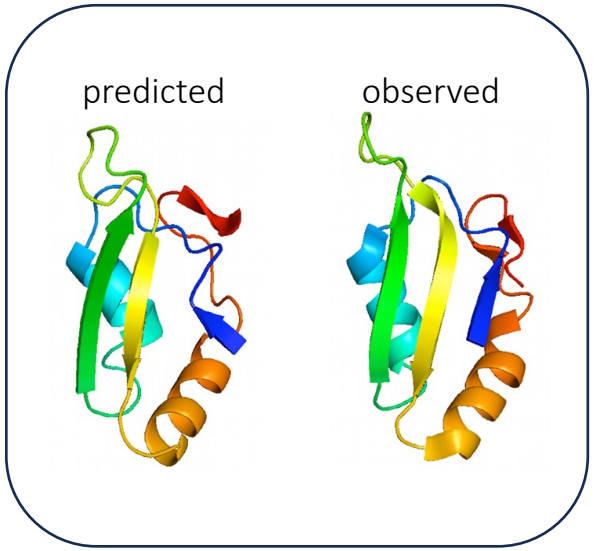
Slide from Ingrahm & Riesselman, MIA 2017

Co-evolving sites (correlated mutations) are likely 3D contacts



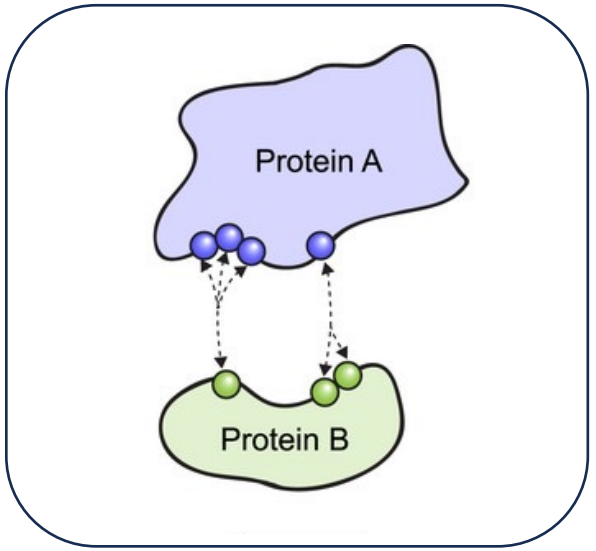
Pairwise interactions model

Structure prediction



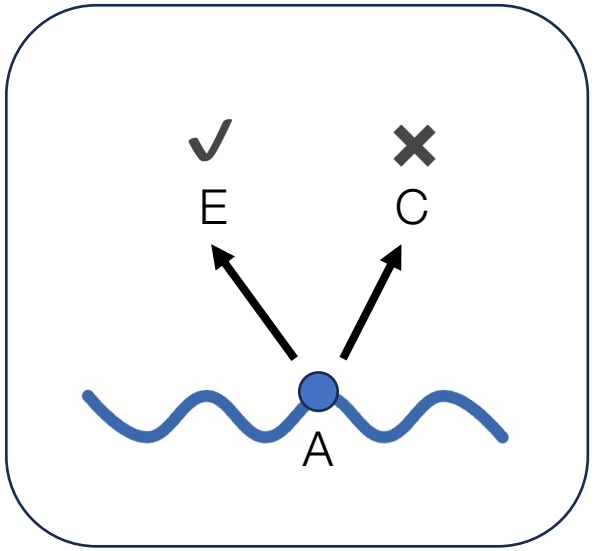
Hopf *et al* 2012 *Cell*
Marks *et al* 2012 *Nature Biotech*

Protein-protein interactions



Green *et al* 2021 *Nature Comm*
Hopf *et al* 2014 *eLife*

Mutation effect prediction

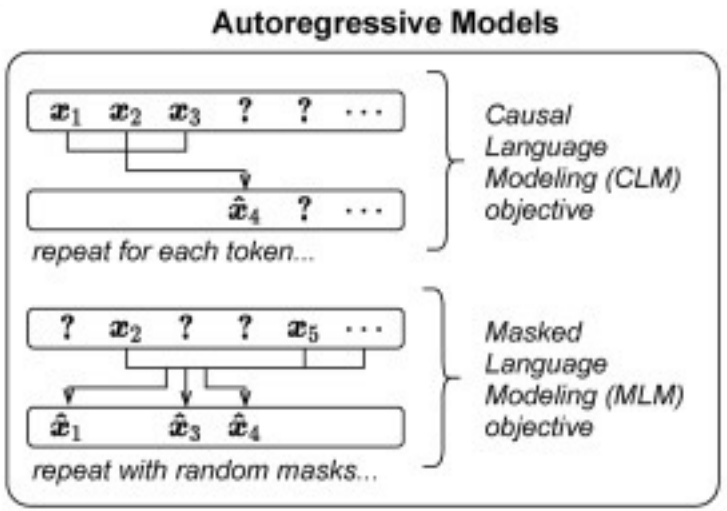


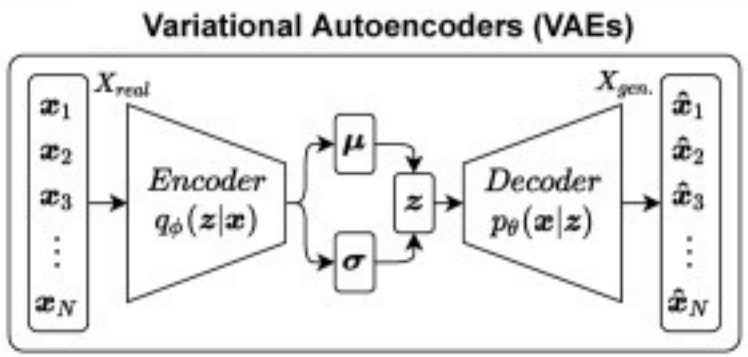
Hopf *et al* 2017 *Nature Biotech*
Riesselman *et al* 2018 *Nature Meth*

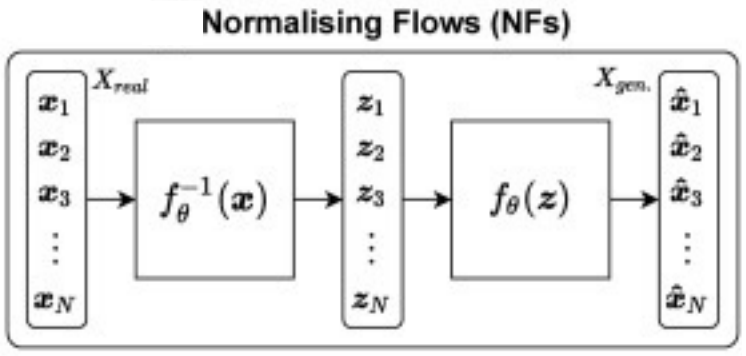
Many ways to model sequences

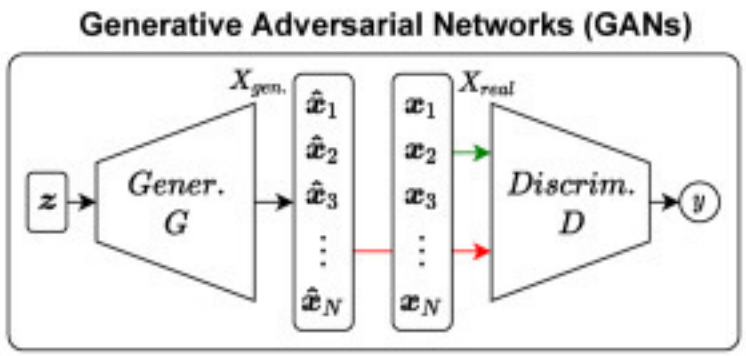
- Site-independent models
- Potts model (pairwise interactions)
- Higher order interactions

Many ways to model higher order interactions









Strokach and Kim 2022, *Curr. Opin. Struct. Biol.*

Protein language models don’t need multiple sequence alignments

Challenges

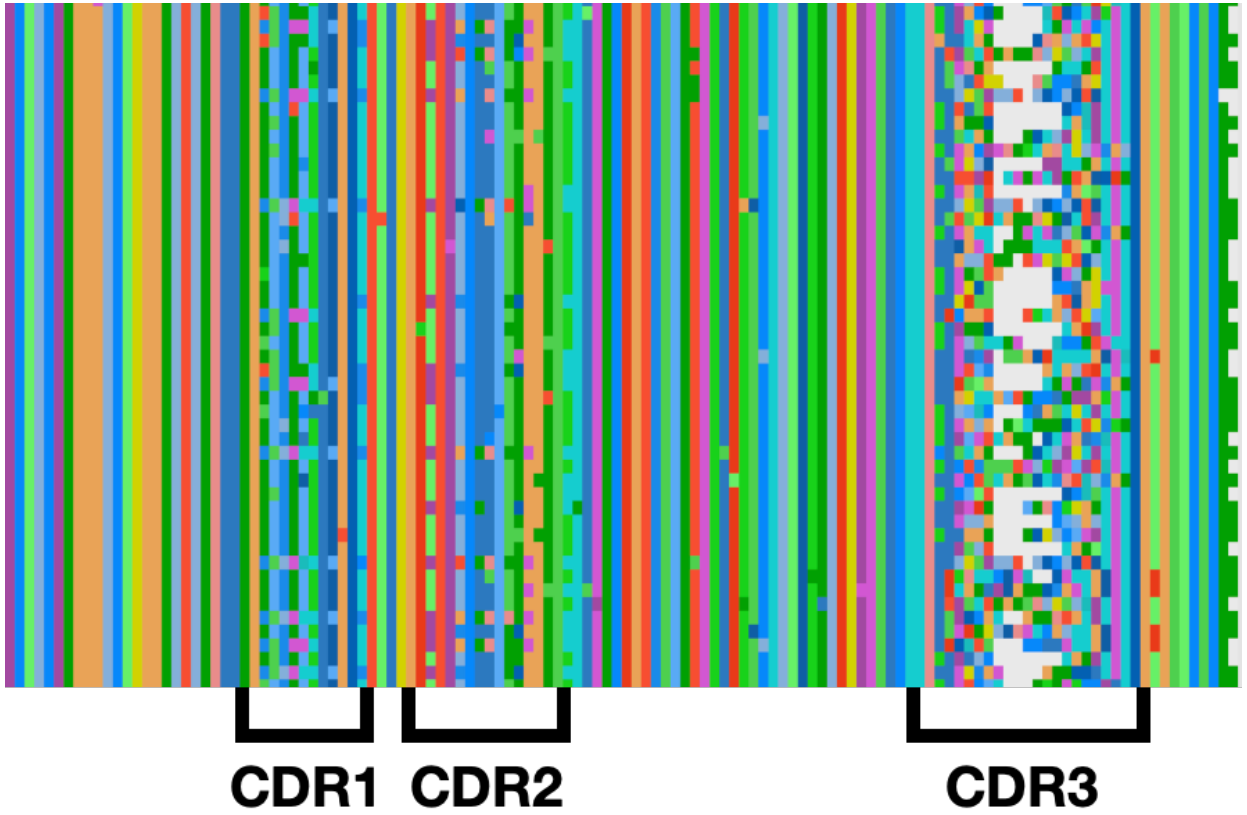
Gap characters

Alignment uncertainty

Insertions and deletions

Slide from Aaron Kollasch

Some proteins like antibodies are just hard to align



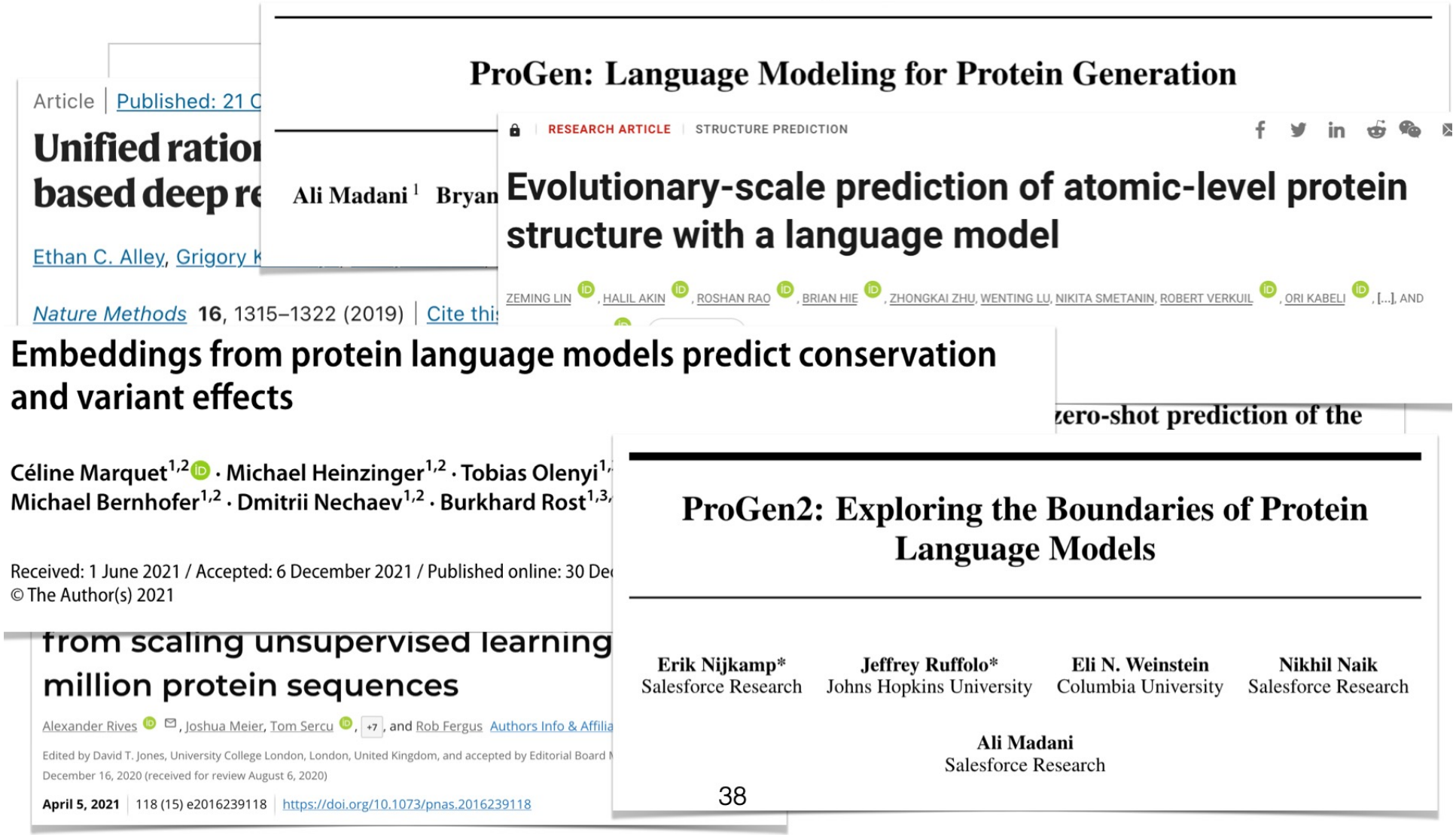
Antibody

Slide from Aaron Kollasch

If there is no alignment...we aren't limited to **one** protein family that aligns together.

What if we built a model of **all** proteins?

Rise of **large language models** for proteins



Slide from Aaron Kollasch

Language modeling training with self-supervision (no labels) and without alignments

Autoregressive (predict next token):

Natural language: The brown fox jumped over the _____

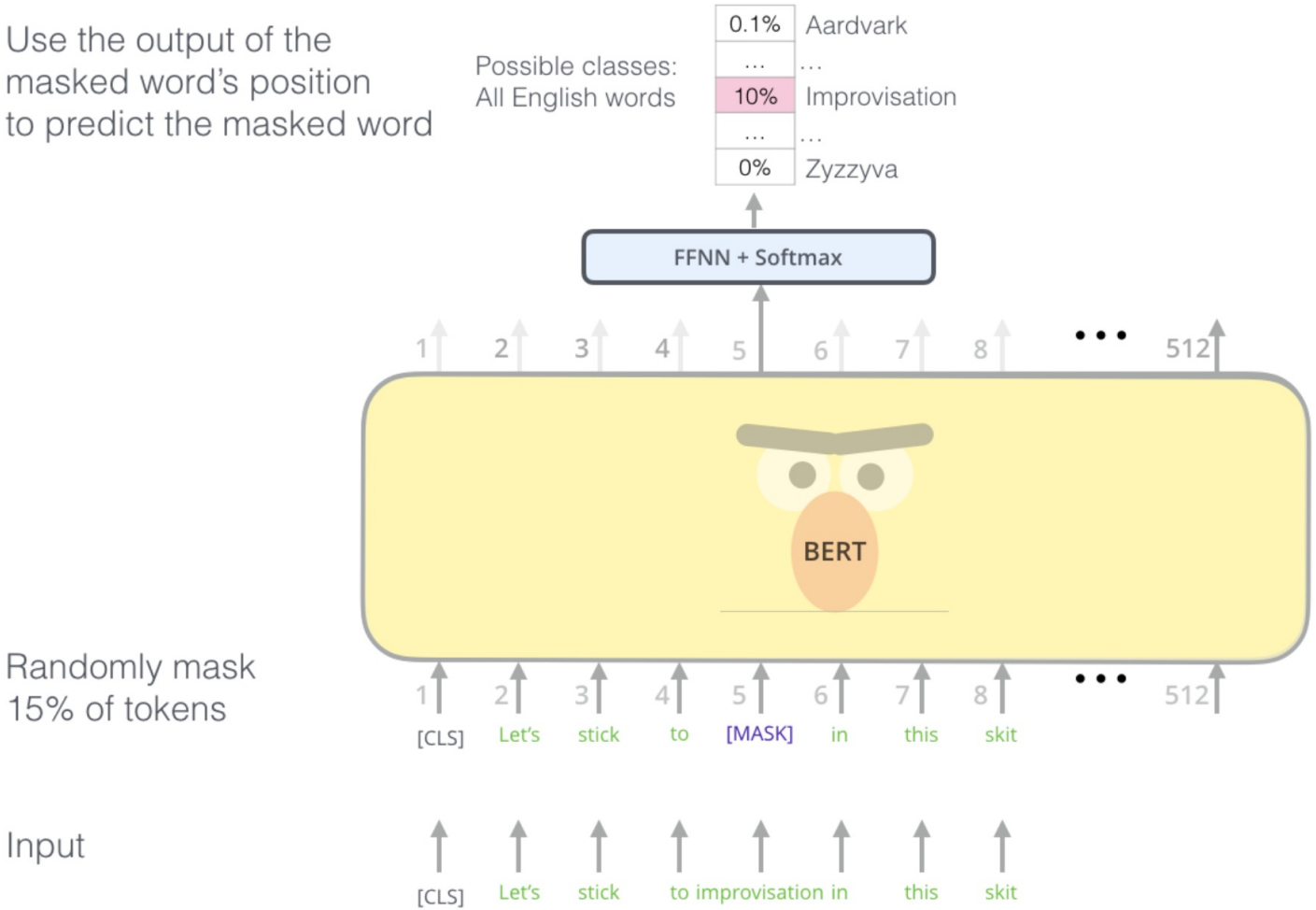
Protein Sequence: GYAMLIVEQGA_

Masked (predict masked tokens):

Natural language: The brown _____ jumped over the ____ dog

Protein Sequence: GYAM_IVEQGAL_SKGV_AIT

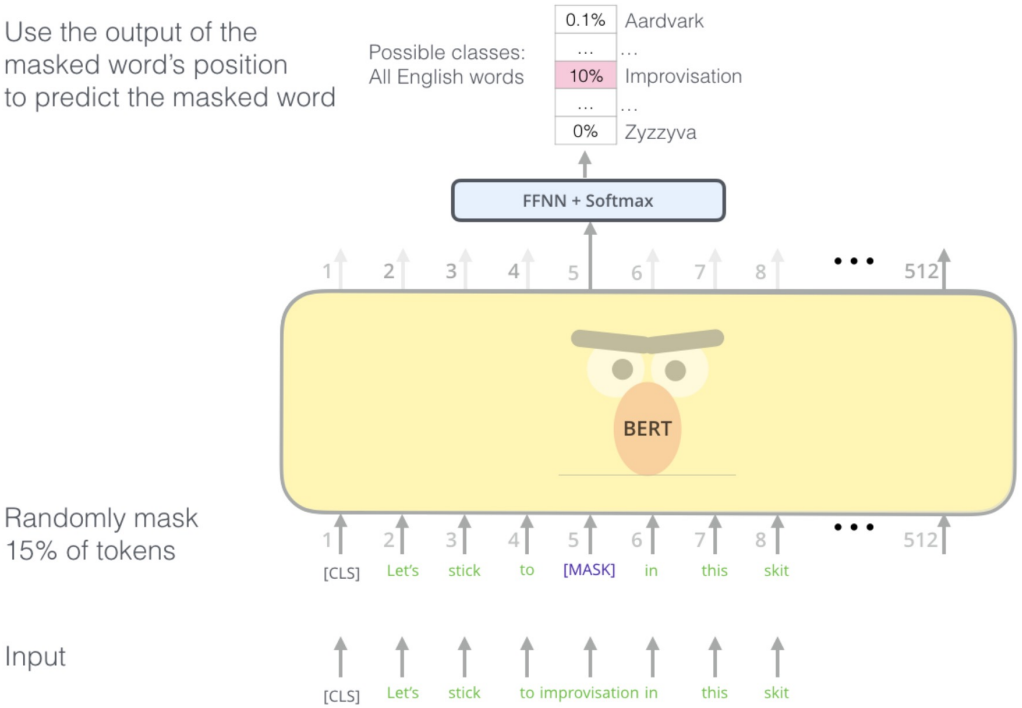
Masked model (pre-)training from BERT



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

https://jalammar.github.io/illustrated-bert/

ESM uses the BERT framework



BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

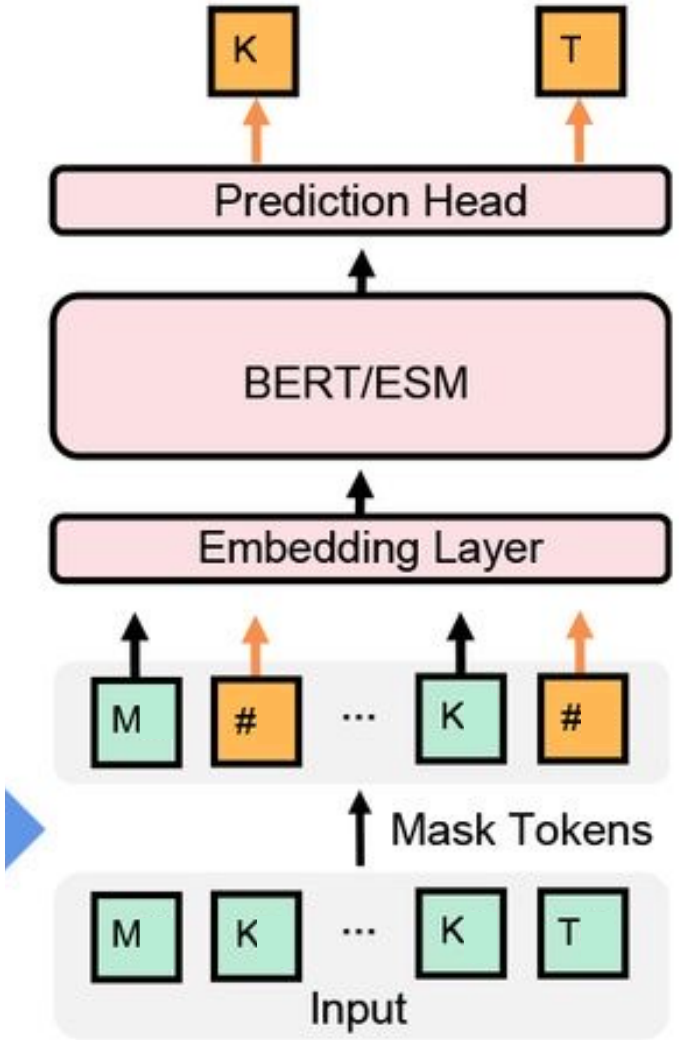
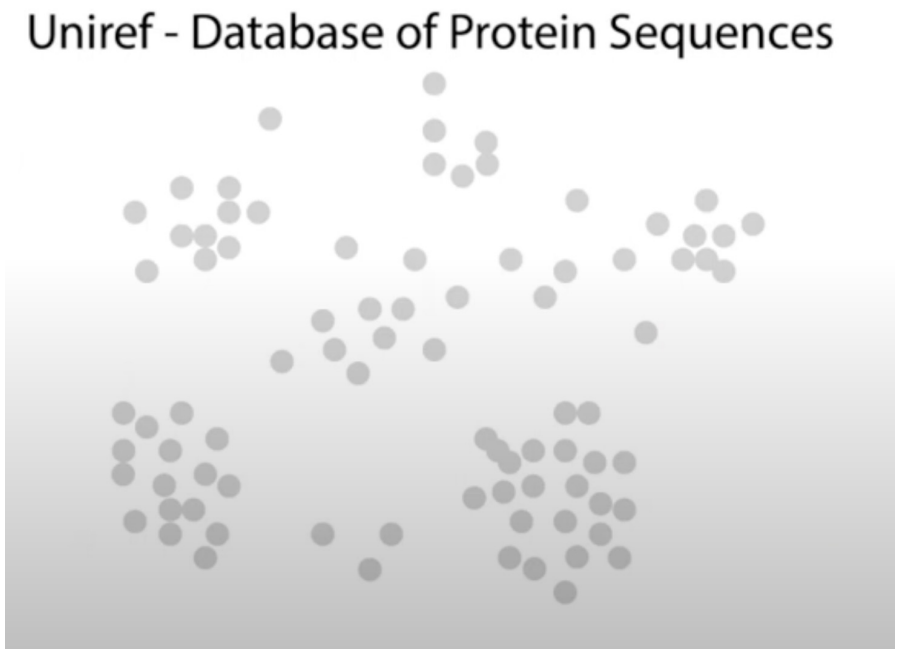


Figure adapted from Su et al. (2024), bioRxiv
https://jalammar.github.io/illustrated-bert/

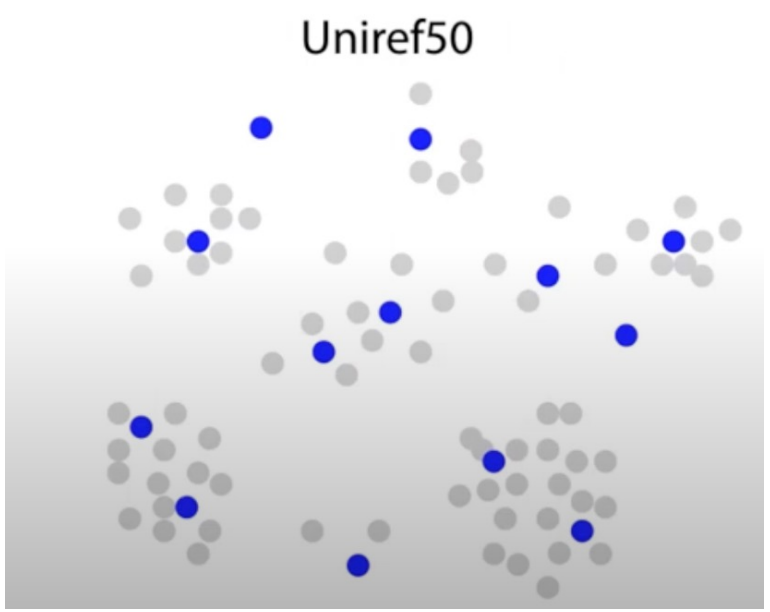
Most protein language models are trained on UniRef



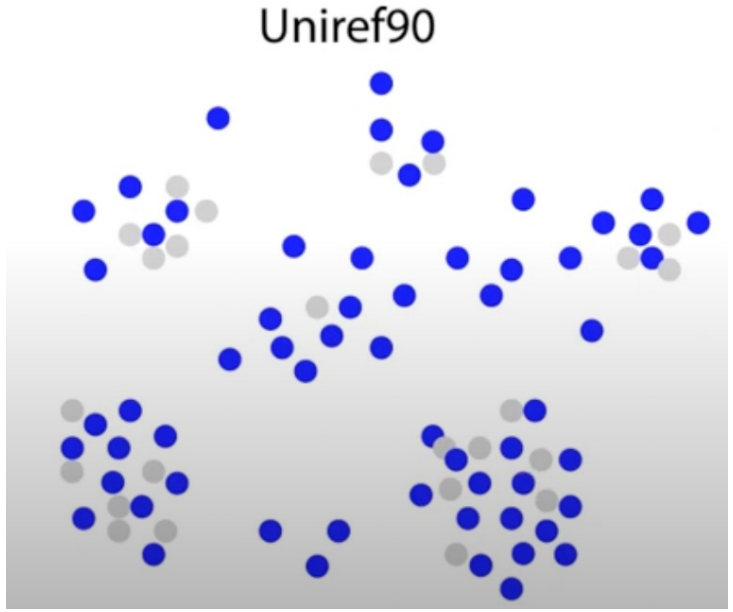
- Protein sequences are not i.i.d.
- Protein databases contain clusters of very similar sequences
- Models trained on all data overfit to largest cluster

Slide adapted from Roshan Rao

Deduplicate and cluster sequences for better performance



50% Sequence Identity

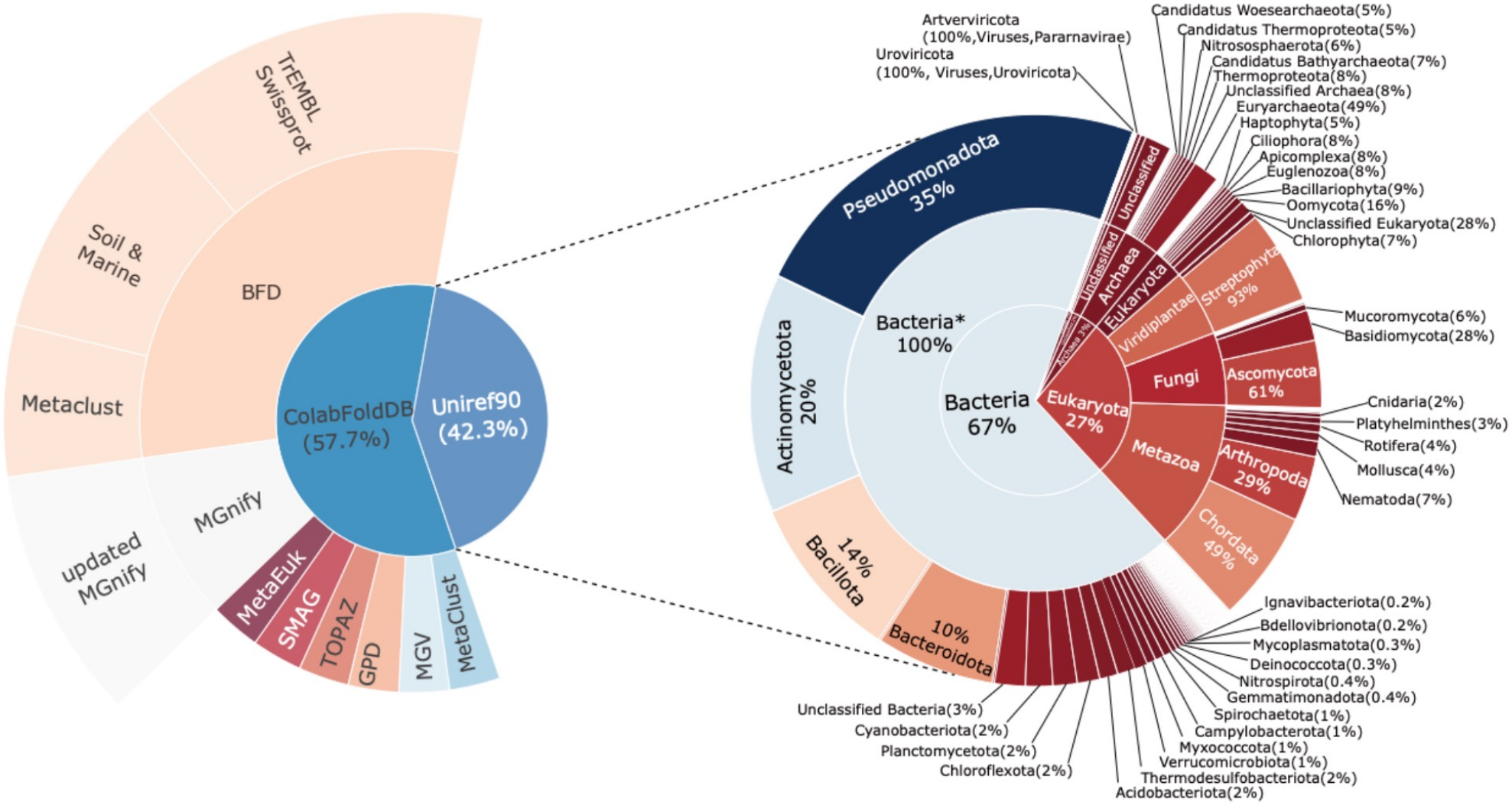


90% Sequence Identity

Depends on the task:

- Uniref50 may be better for **structure prediction**
- Greater number of sequences in Uniref90 is better for **mutation effect prediction**

Slide adapted from Roshan Rao



Chen et al. BioRxiv

Most protein language models are trained on UniRef

Model and Repository	Input	Network	#Embedding	#Param.	Pretext Task	Pre-training Dataset
PRoBERTa [158]	Seq	Transformer	768	44M	Masked Language Modeling	Swiss-Prot
ESM-1b [14]	Seq	Transformer	1280	650M	Masked Language Modeling	UniParc
ProtTXL [12]	Seq	Transformer-XL	1024	562M	Masked Language Modeling	BFD100, UniRef100
ProtBert [12]	Seq	BERT	1024	420M	Masked Language Modeling	BFD100, UniRef100
ProtXLNet [12]	Seq	XLNet	1024	409M	Masked Language Modeling	UniRef100
ProtAlbert [12]	Seq	ALBERT	4096	224M	Masked Language Modeling	UniRef100
ProtElectra [12]	Seq	ELECTRA	1024	420M	Masked Language Modeling	UniRef100
ProtT5 [12]	Seq	T5	1024	11B	Masked Language Modeling	UniRef50, BFD100
PMLM [159]	Seq	Transformer	1280	715M	Masked Language Modeling	UniRef50
MSA Transformer [160]	MSA	Transformer	768	100M	Masked Language Modeling	UniRef50, UniClust30
ProteinLM [161]	Seq	BERT	-	3B	Masked Language Modeling	Pfam
PLUS-RNN [162]	Seq	RNN	2024	59M	Masked Language Modeling Same-Family Prediction	Pfam
CARP [163]	Seq	CNN	1280	640M	Masked Language Modeling	UniRef50
AminoBERT [22]	Seq	Transformer	3072	-	Masked Language Modeling	UniParc
OmegaPLM [164]	Seq	GAU [165]	1280	670M	Masked Language Modeling Span and Sequential Masking	UniRef50
ProGen2 [166]	Seq	Transformer	4096	6.4B	Masked Language Modeling	UniRef90, BFD30, BFD90
ProtGPT2 [167]	Seq	GPT-2 [168]	1280	738M	Next Amino Acid Prediction	UniRef50
RITA [169]	Seq	GPT-3 [55]	2048	1.2B	Next Amino Acid Prediction	UniRef100
ESM-2 [13]	Seq	Transformer	5120	15B	Masked Language Modeling	UniRef50

Hu et al.

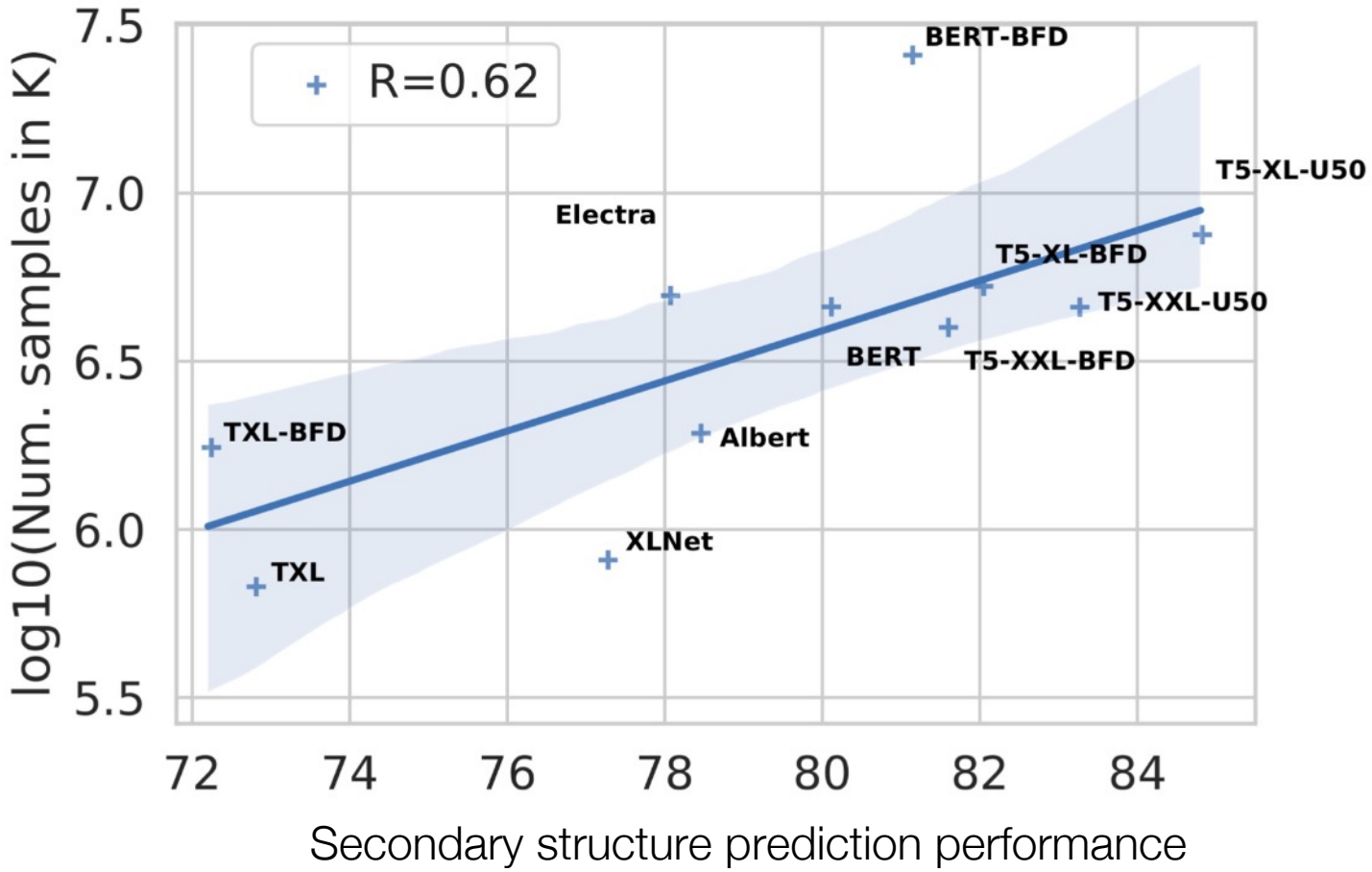
What are the sizes of these datasets?

Dataset	#Proteins	Disk Space	Description
UniProtKB/Swiss-Prot [9]	500K	0.59GB	knowledgebase
UniProtKB/TrEMBL [9]	229M	146GB	knowledgebase
UniRef100 [215]	314M	76.9GB	clustered sets of sequences
UniRef90 [215]	150M	34GB	90% identity
UniRef50 [215]	53M	10.3GB	50% identity
UniParc [9]	528M	106GB	sequence
PDB [8]	180K	50GB	3D structure
CATH4.3 [216]	-	1073MB	hierarchical classification
BFD [217]	2500M	272GB	sequence profile
Pfam [218]	47M	14.1GB	protein families
AlphaFoldDB [219]	214M	23 TB	predicted 3D structures
ESM Metagenomic Atlas [13]	772M	-	predicted metagenomic protein structures
ColAbFoldDB [170]	950M	-	an amalgamation of various metagenomic databases
ProteinKG25 [10]	5.6M	147MB	a knowledge graph dataset with GO
Uniclust30 [220]	-	6.6GB	clustered protein sequences
SCOP [221]	-	-	structural classification
SCOPe [222]	-	86MB	an extended version of SCOP
OpenProteinSet [223]	16M	-	MSAs

BFD is the biggest (and most fantastic)

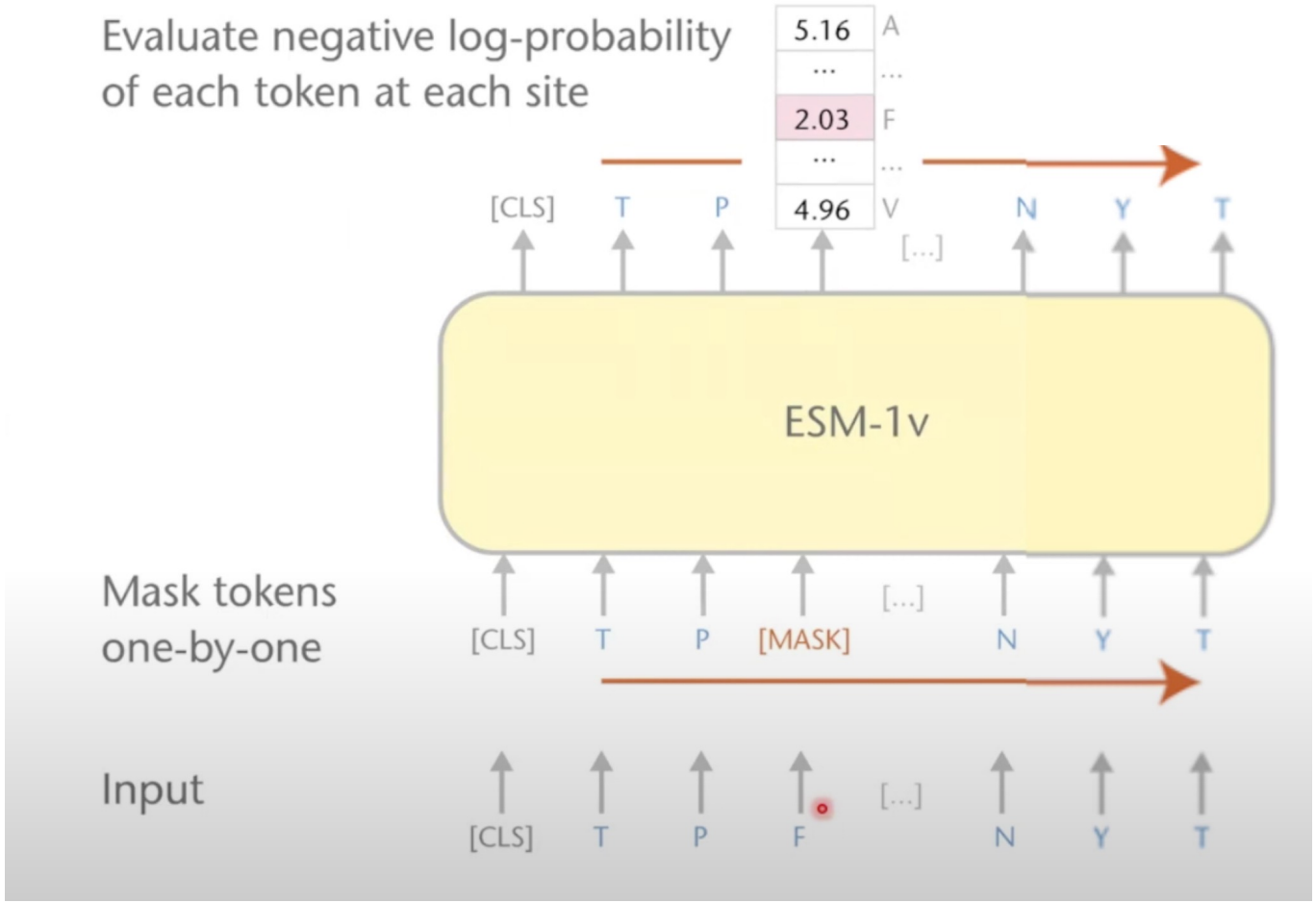
Flamholz et al.

Model performance improves with more training sequences



Elnaggar et al. (2022) IEEE

Mutation effects from a protein language model



Compute **Log Ratio** for each mutant

$$\log \frac{p(\mathbf{x}_{\text{mut}}|\boldsymbol{\theta})}{p(\mathbf{x}_{\text{wild}}|\boldsymbol{\theta})}$$

Meier et al. (2022) Neurips
Slide from Roshan Rao

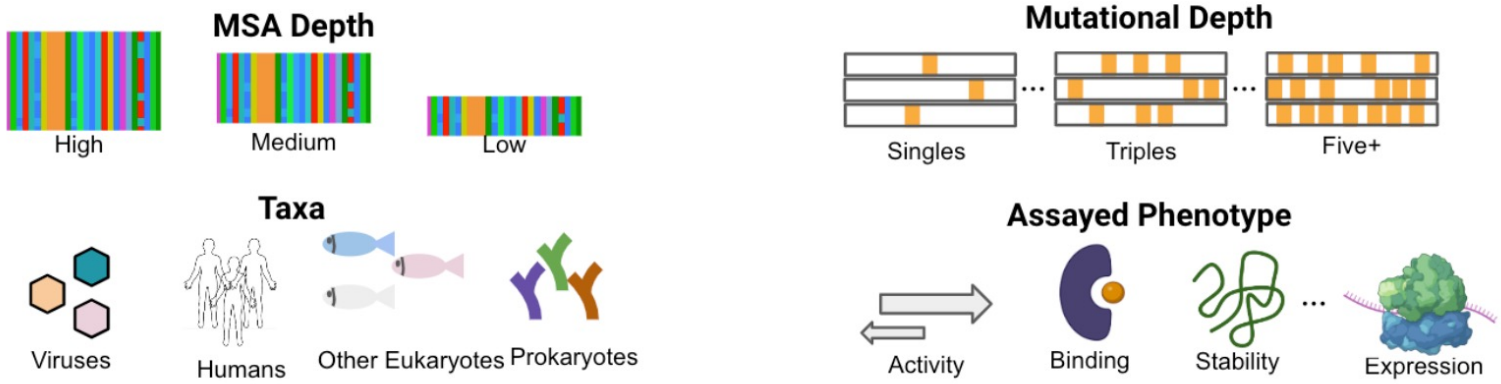
Common evaluation tasks

- **Contact Prediction**: predict whether two residues are in the pre-defined proximity
- **Fold Prediction**: assign sequence to one to 1,195 known folds
- **Secondary Structure Prediction**: classify sequence into Helix, Strand, Coil
- **Solubility**: binary
- **Stability:** whether protein remains folded
- **Localization:** distribution into 100 Unique Subcellular catergories
- **TCR-pMHC Affinity:** interaction between T cell receptors and peptide-major hits compatibility complex
- ….

How to evaluate mutation effect prediction across all of the models?

ProteinGym covers a diverse array of proteins

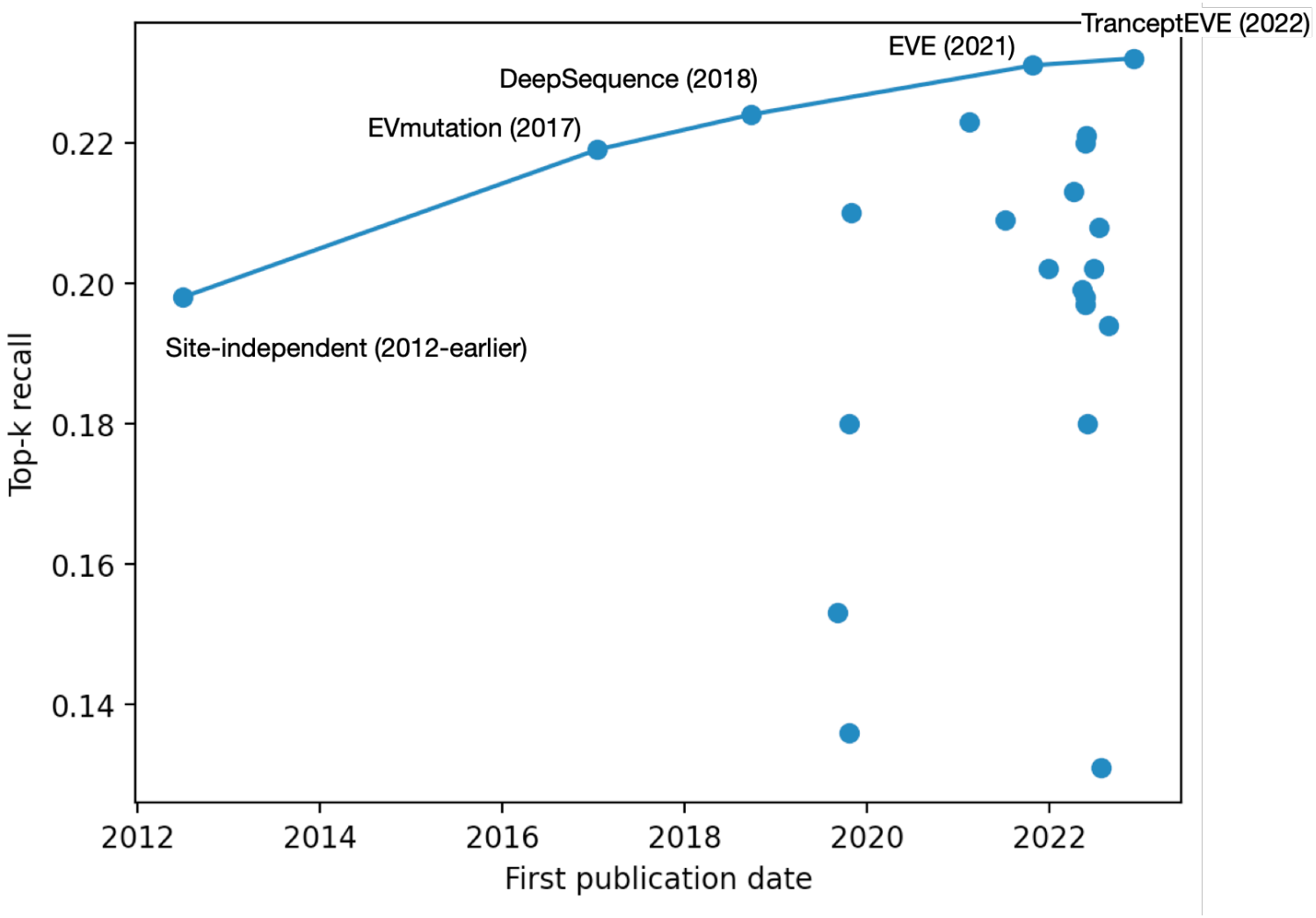
~30 proteins (2018) ➔ >200 proteins (now)



Total: 2.8 million variants

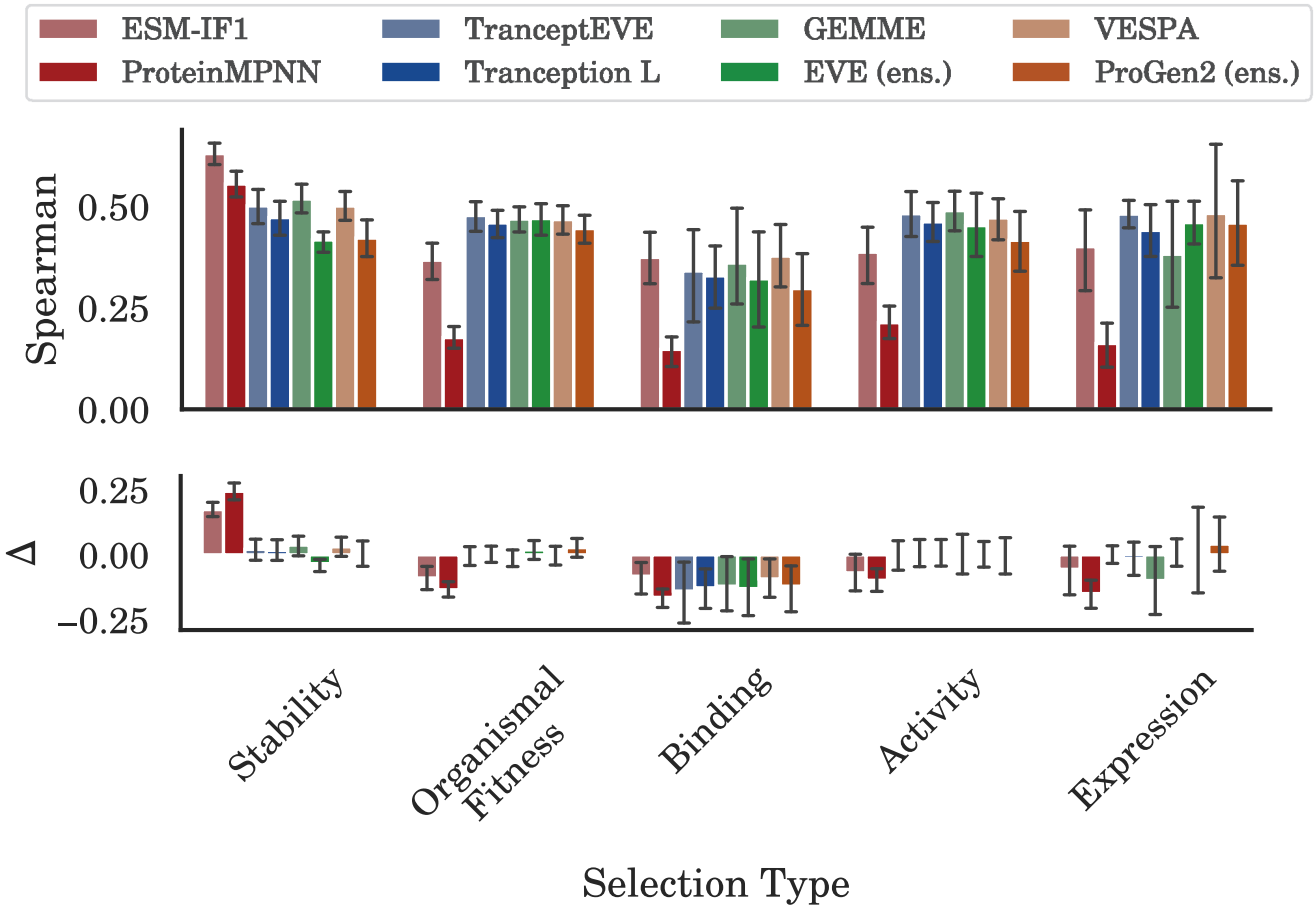
Notin et al. (2023)
Slide from Aaron Kollasch

Performance against Deep Mutational Scans has improved over time



Notin et al. (2023)
Slide from Aaron Kollasch

Different models better at predicting different functions



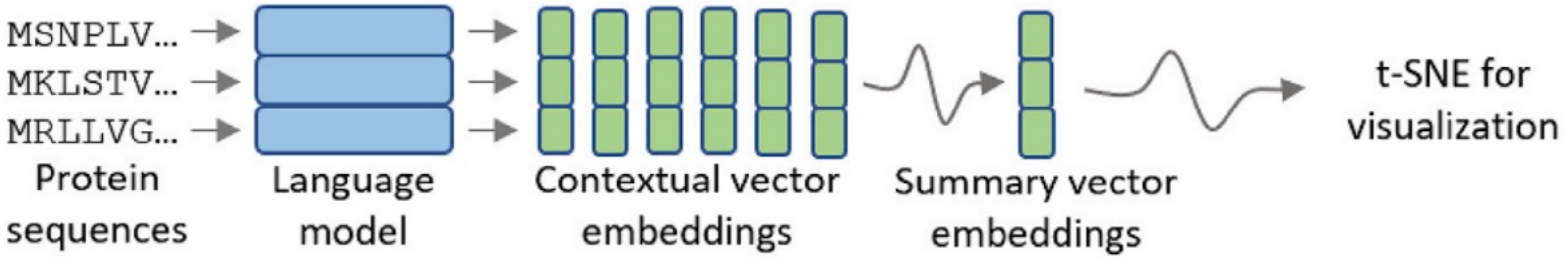
Notin et al. (2023)
Slide from Aaron Kollasch

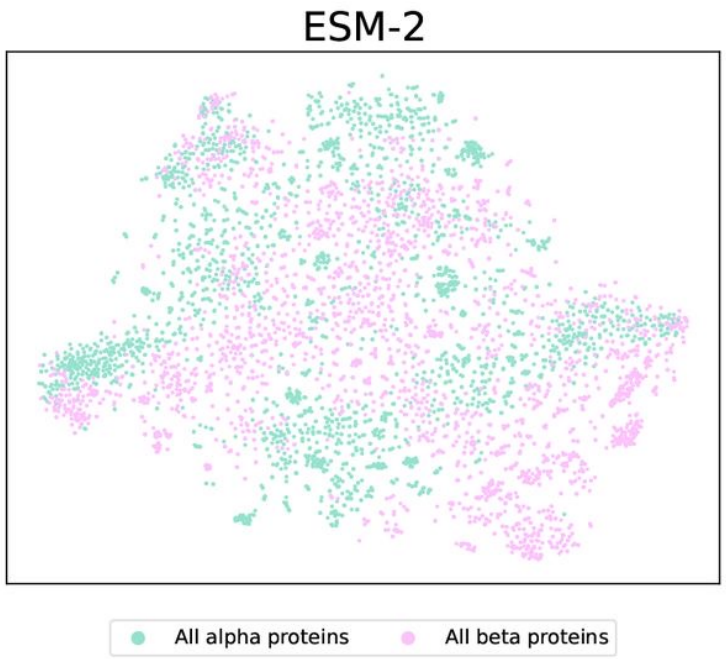
Models do better with more relevant sequences

Model type	Model name	Spearman by MSA depth (↑)			
		Low	Medium	High	All
Alignment-based	Site-Independent	0.426	0.373	0.320	0.373
	WaveNet	0.207	0.255	0.207	0.223
	EVmutation	0.403	0.423	0.410	0.412
	DeepSequence (ens.)	0.383	0.428	0.473	0.428
	EVE (ens.)	0.425	0.453	0.481	0.453
	GEMME	**0.455**	**0.470**	0.497	**0.474**
Protein language	UniRep	0.181	0.161	0.209	0.184
	CARP (640M)	0.314	0.375	0.428	0.372
	ESM-1b	0.350	0.398	0.482	0.410
	ESM-2 (15B)	0.357	0.414	0.473	0.415
	RITA XL	0.315	0.382	0.412	0.370
	ESM-1v (ens.)	0.326	0.418	0.502	0.415
	ProGen2 XL	0.354	0.405	0.444	0.401
	VESPA	0.427	0.455	0.484	0.455
Hybrid	UniRep evotuned	0.330	0.344	0.372	0.349
	MSA Transformer (ens.)	0.404	0.450	0.488	0.447
	Tranception L	0.432	0.438	0.473	0.448
	TranceptEVE L	0.451	0.467	0.492	0.470
Inverse Folding	ESM-IF1	0.300	0.431	**0.544**	0.425
	MIF-ST	0.376	0.403	0.456	0.412
	ProteinMPNN	0.173	0.280	0.434	0.296

Notin et al. (2023)
Slide from Aaron Kollasch

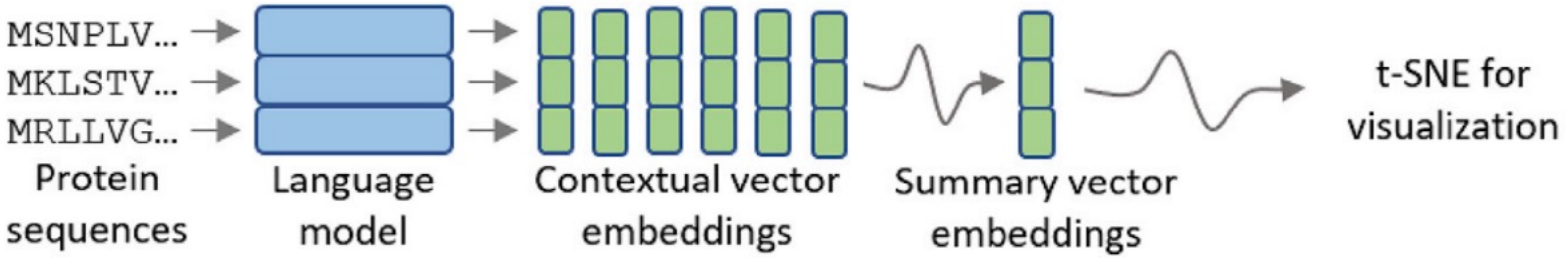
ESM embedding space

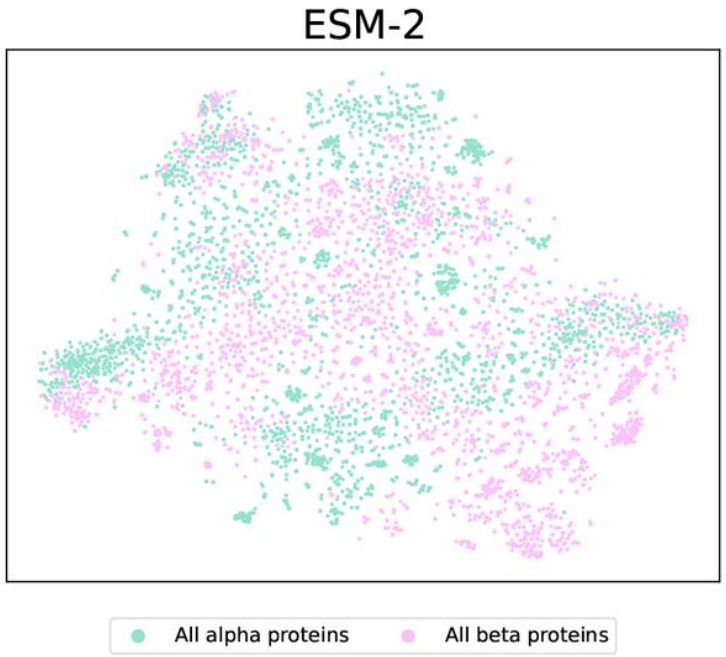


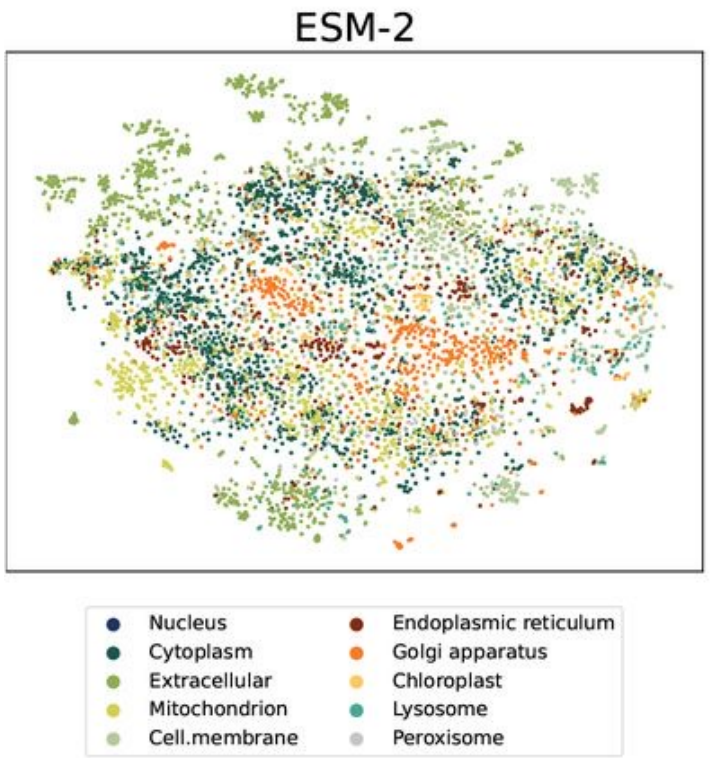


Su et al. (2024), BioRxiv
Bepler et al. (2021), Cell Systems

ESM embedding space





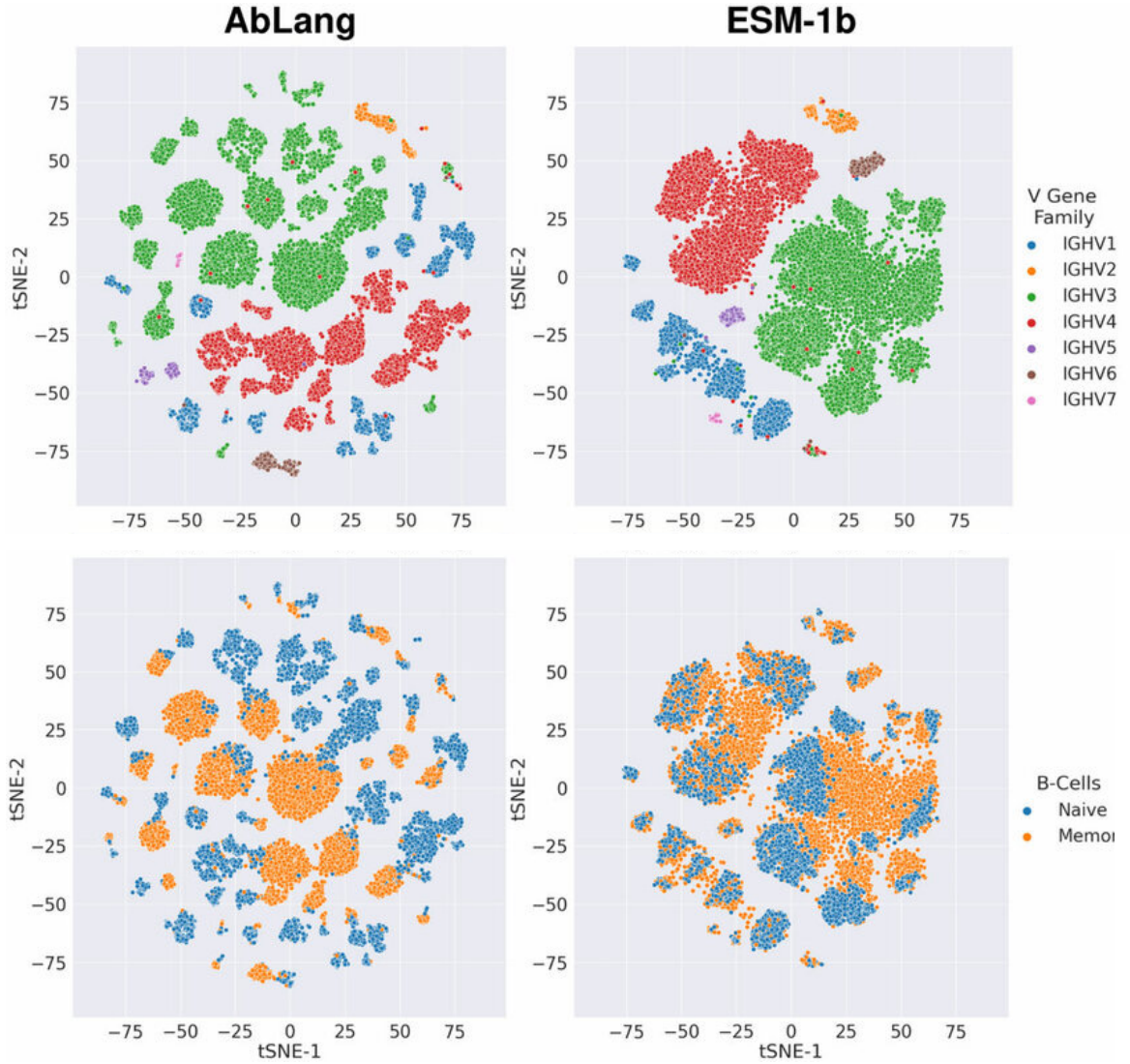


Su et al. (2024), BioRxiv
Bepler et al. (2021), Cell Systems

Structure-aware tokens separate embeddings by structure type

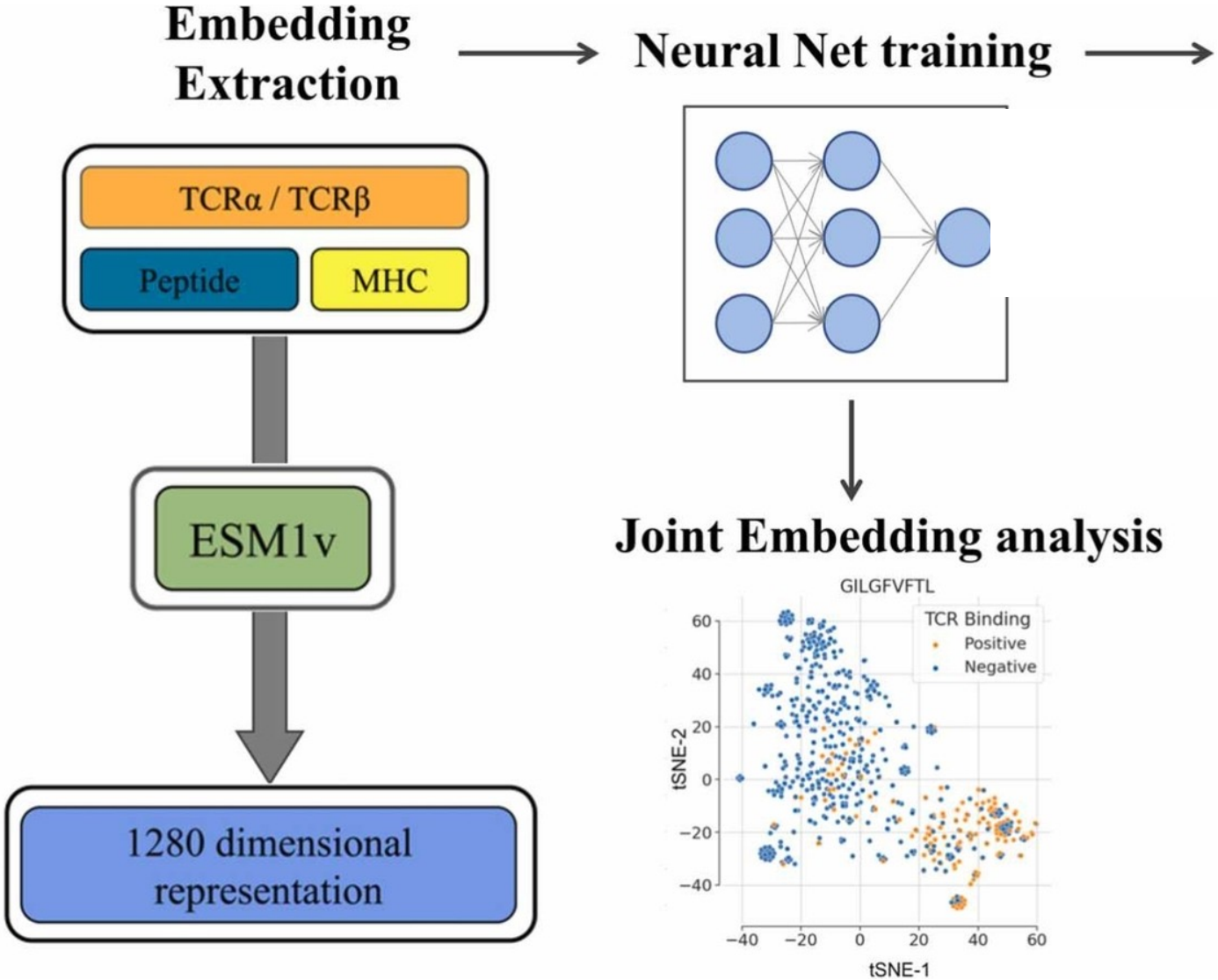
Su et al. (2024), BioRxiv

Focusing on one class of proteins (i.e., antibodies) leads to informative embeddings for that class



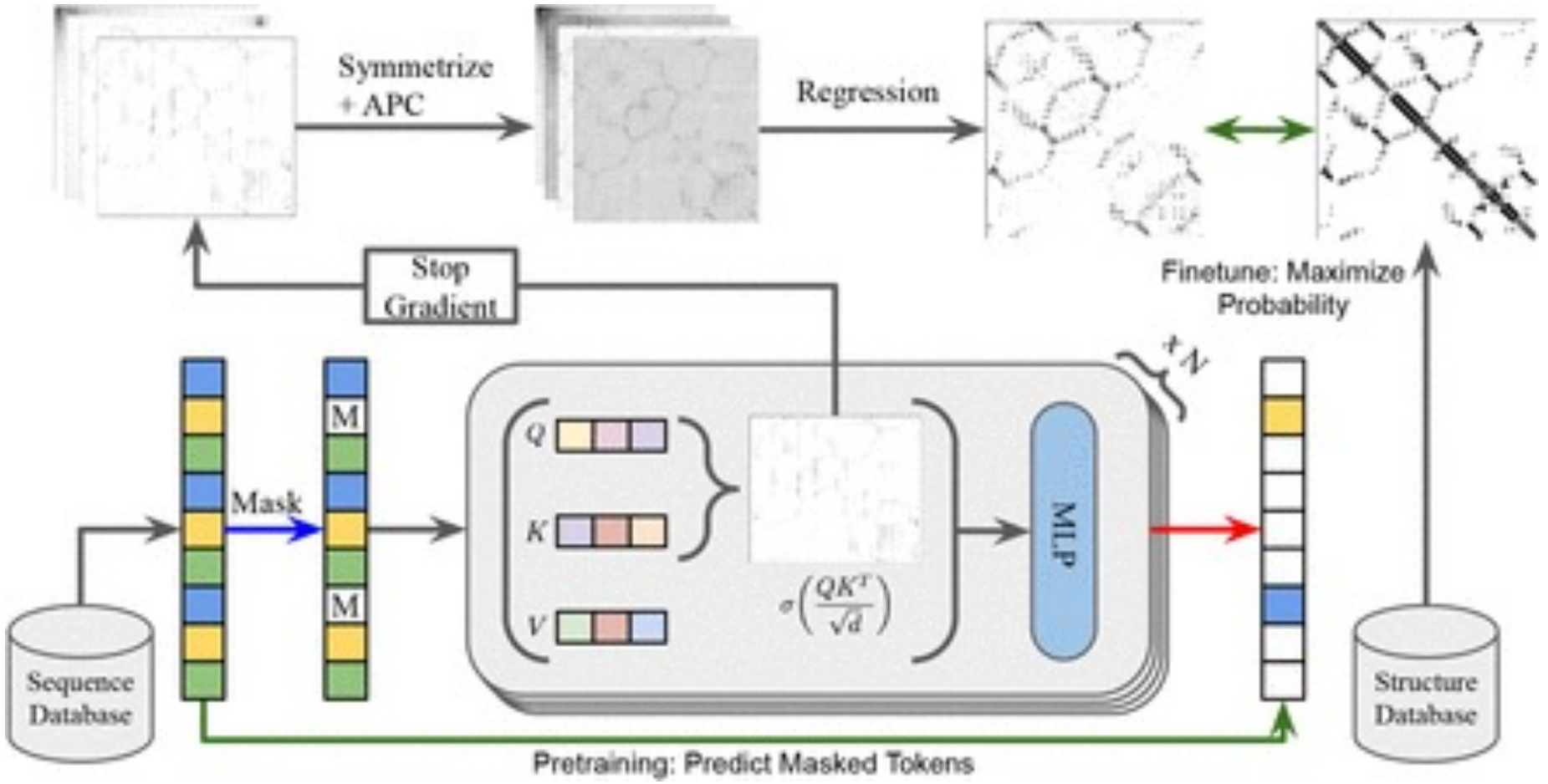
Olson et al. (2022), BioRxiv

Embeddings as inputs into supervised models



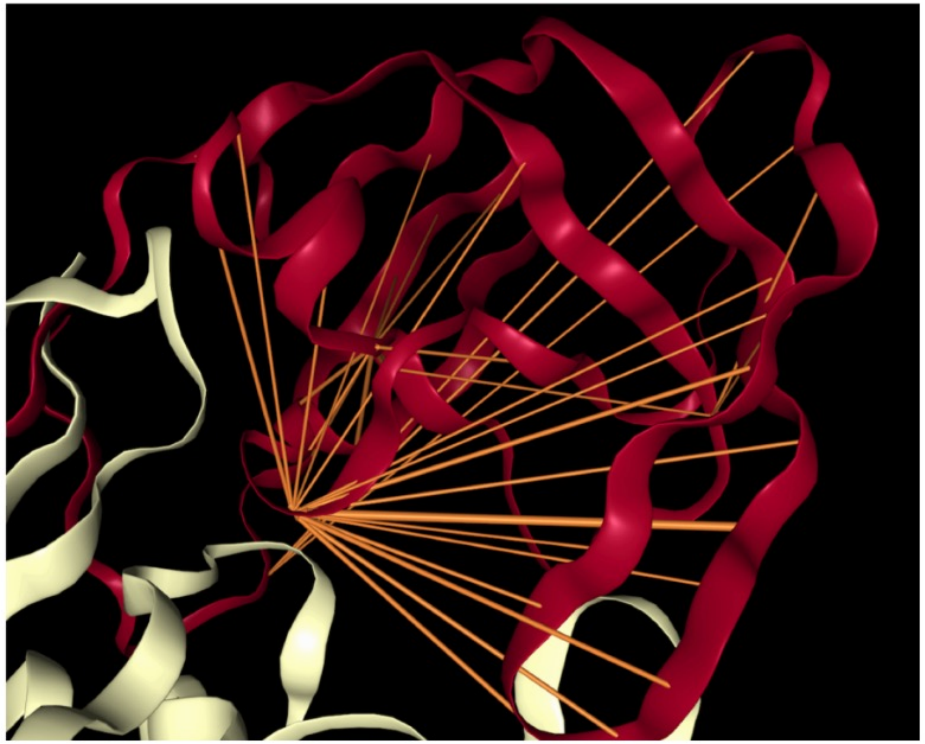
Yadav et al. (2024)

Language model attention maps learn structural contacts



Rao et al. (2020)

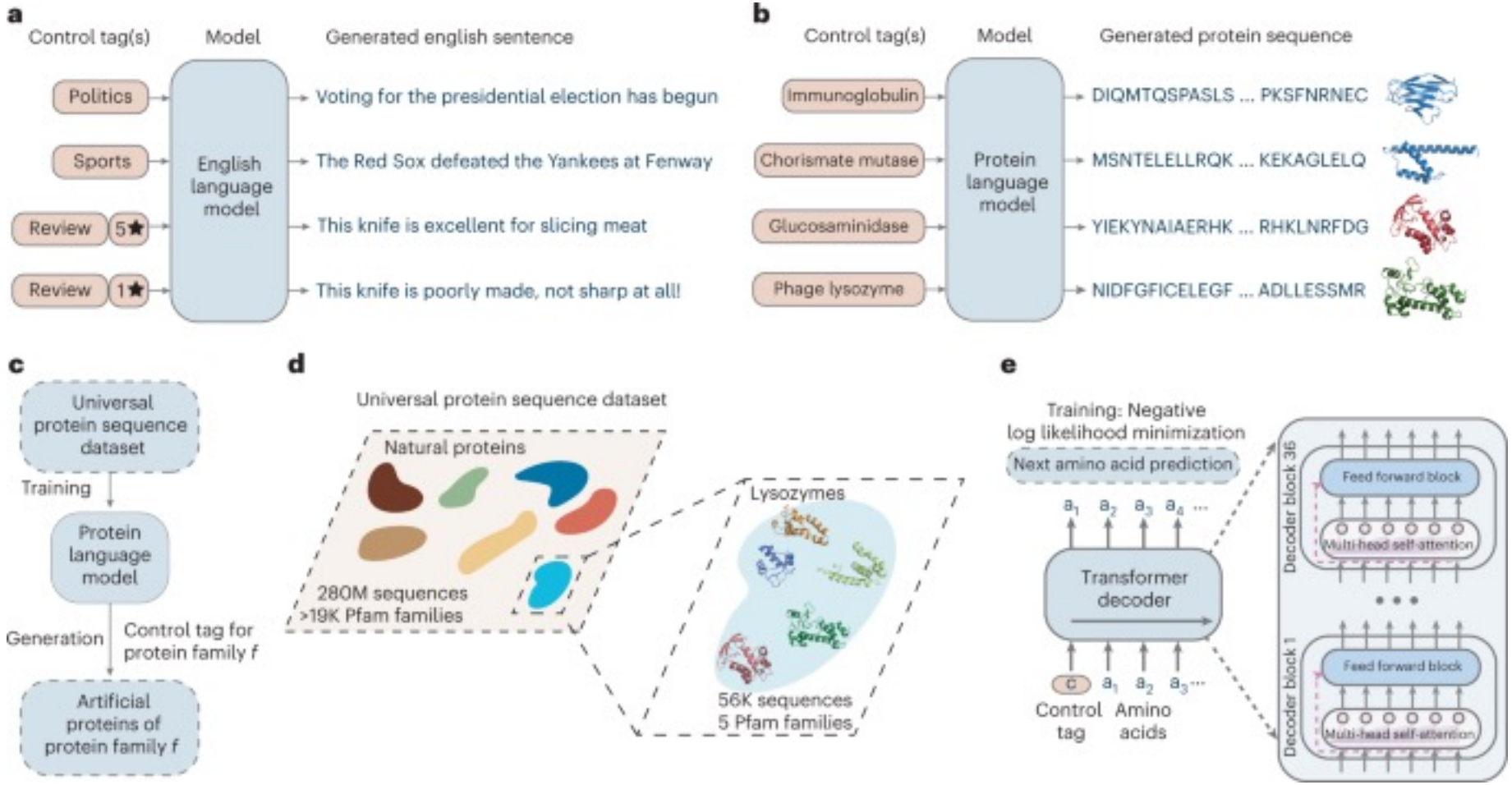
(a) Attention in head 12-4, which targets amino acid pairs that are close in physical space (see inset subsequence 117D-157I) but lie apart in the sequence. Example is a *de novo* designed TIM-barrel (5BVL) with characteristic symmetry.



(b) Attention in head 7-1, which targets binding sites, a key functional component of proteins. Example is HIV-1 protease (7HVP). The primary location receiving attention is 27G, a binding site for protease inhibitor small-molecule drugs.

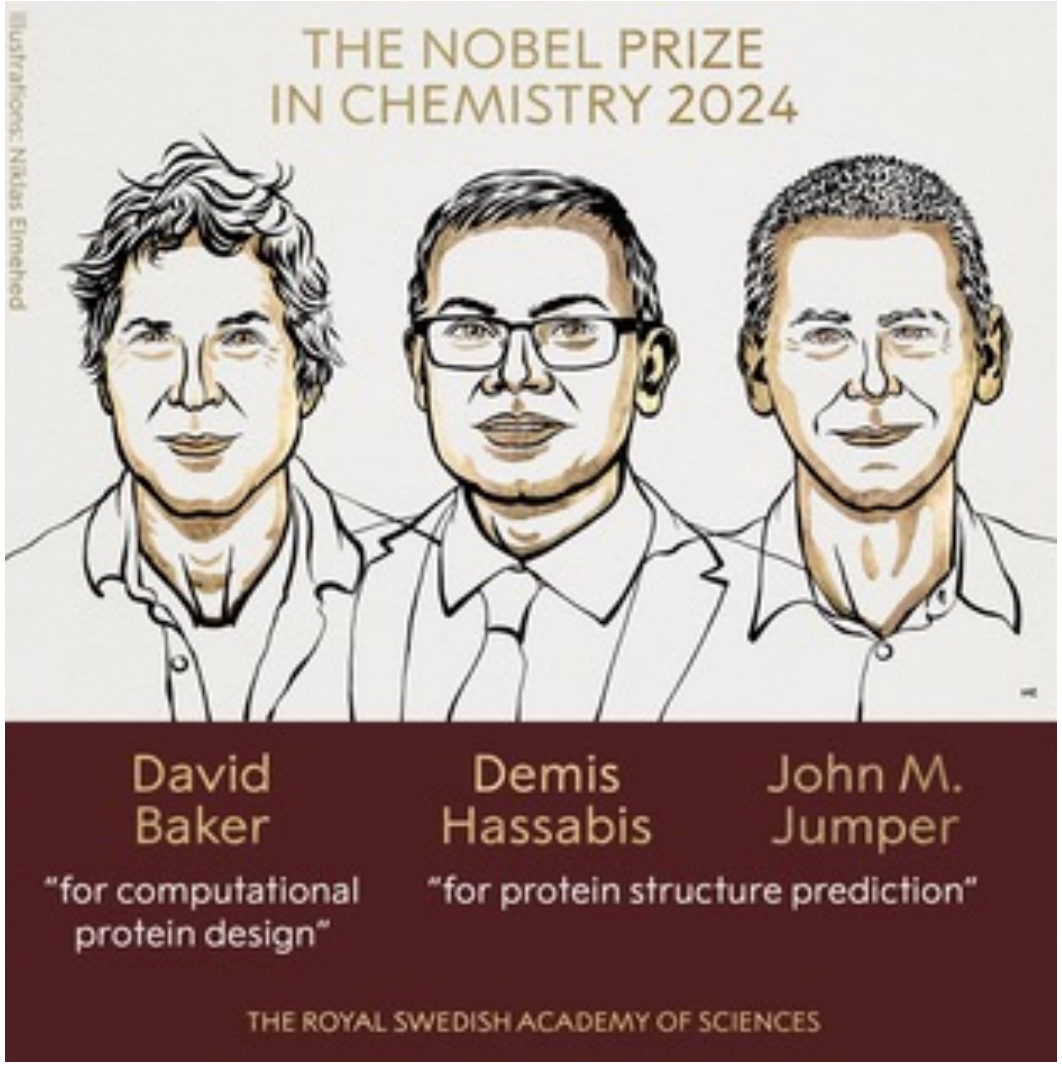
https://openreview.net/pdf?id=YWtLZvLmud7

Protein language models can design new proteins!



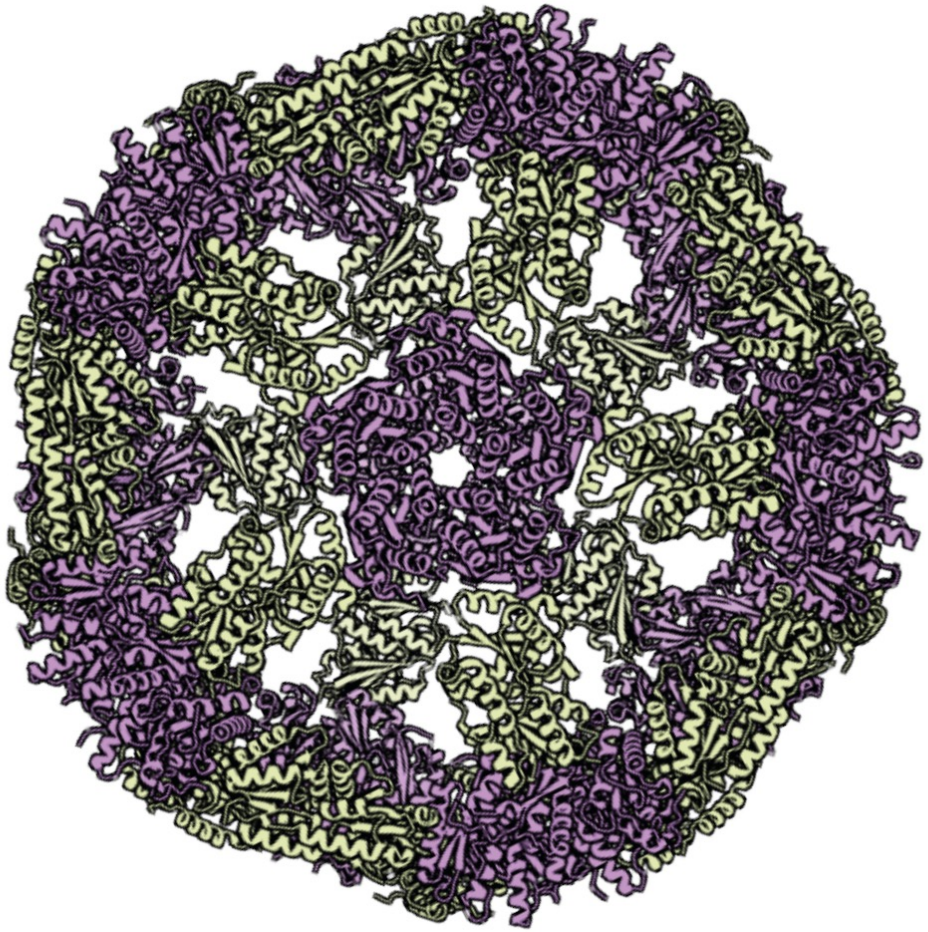
Madani et al. (2023) Nature Biotech.

An exciting time for computational protein design!



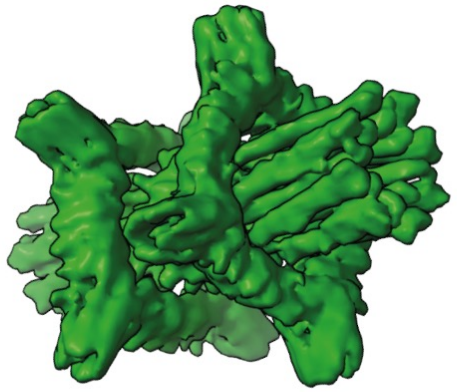
Nobelprize.org

Figure 4. Proteins developed using Baker's program Rosetta.

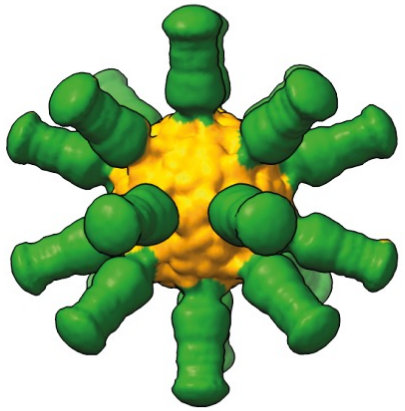


2016: New nanomaterials where up to 120 proteins spontaneously link together.

2017: Proteins that bind to an opioid called fentanyl (purple). These could be used to detect fentanyl in the environment.



2022: Proteins that function as a type of molecular rotor.

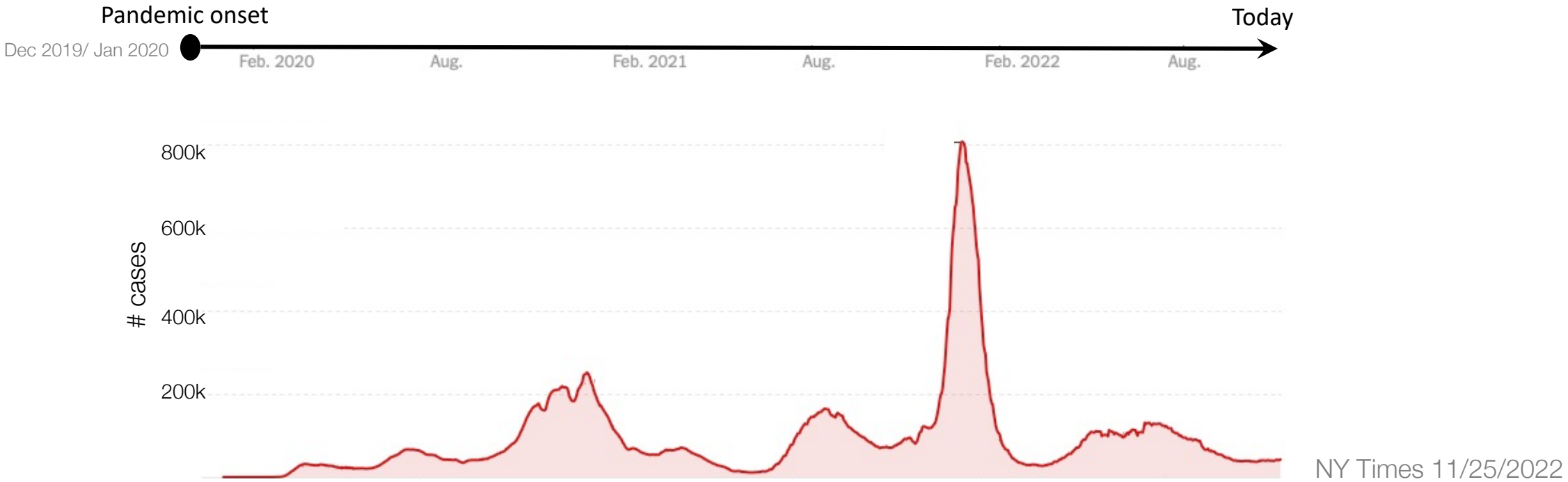


2021: Nanoparticles (yellow) with proteins imitating influenza virus on the surface (green) that can be used as a vaccine for influenza. Successful in animal models.

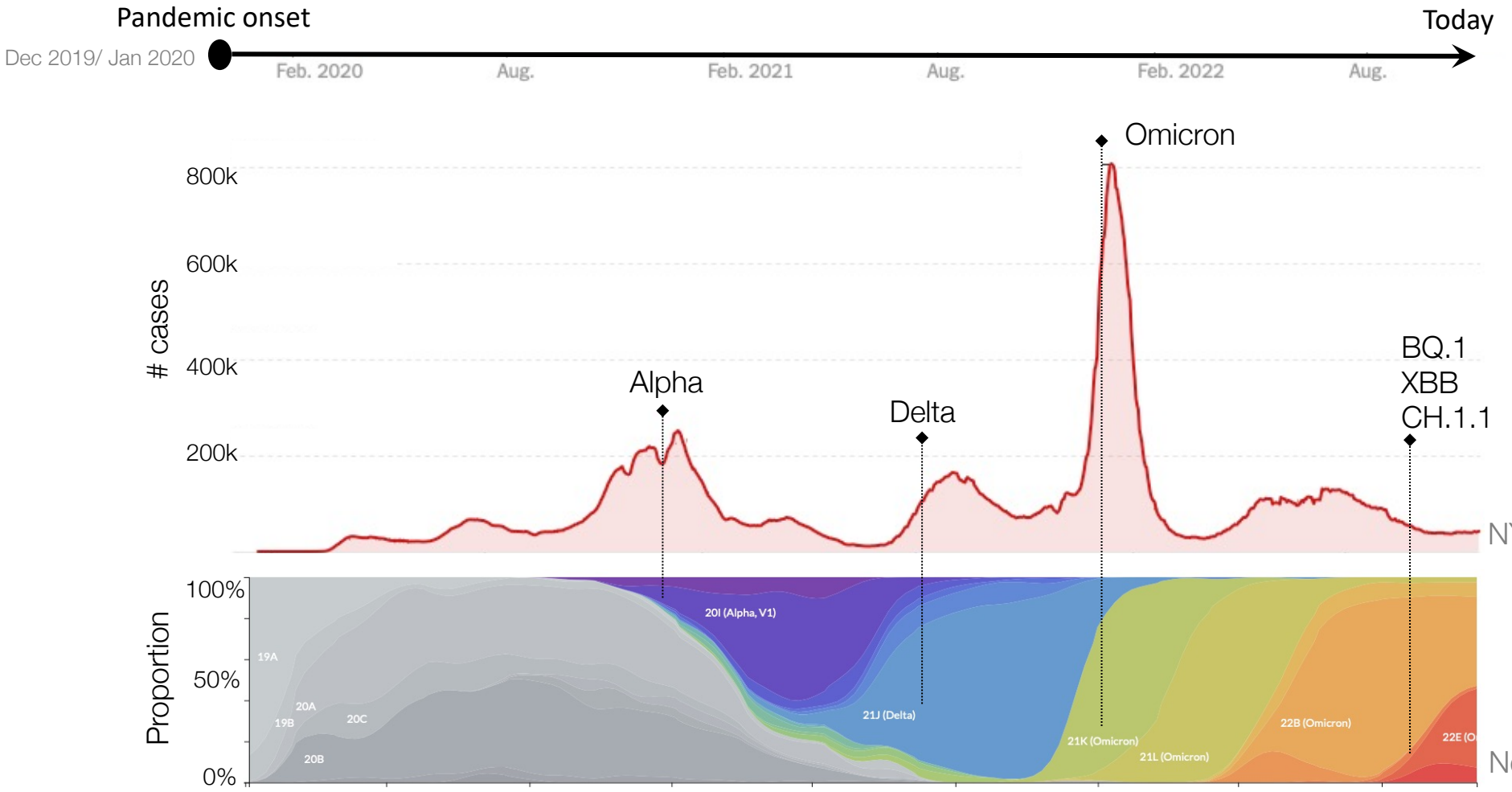
2024: Geometrically shaped proteins that can change their shape due to external influences. Could be used for producing tiny sensors.

Nobelprize.org

A case study on predicting viral evolution of the SARS-CoV-2 Spike protein



NY Times 11/25/2022



NY Times 11/25/2022

Nextstrain.org 11/25/2022

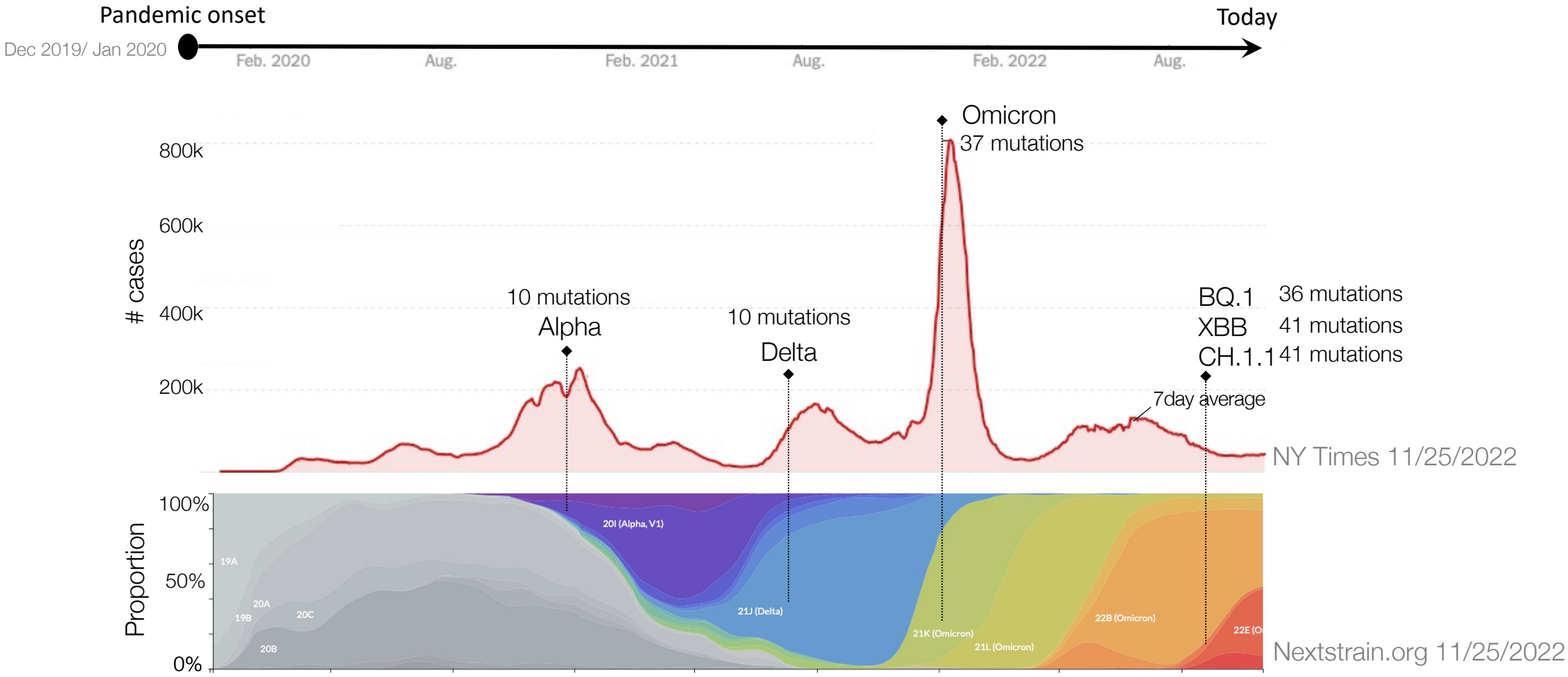
HEALTH

The coronavirus isn't mutating quickly, suggesting a vaccine would offer lasting protection

By Joel Achenbach

March 24, 2020 at 4:30 p.m. EDT

Washington post – March 2020



HEALTH

The coronavirus isn't mutating quickly, suggesting a vaccine would offer lasting protection

By Joel Achenbach

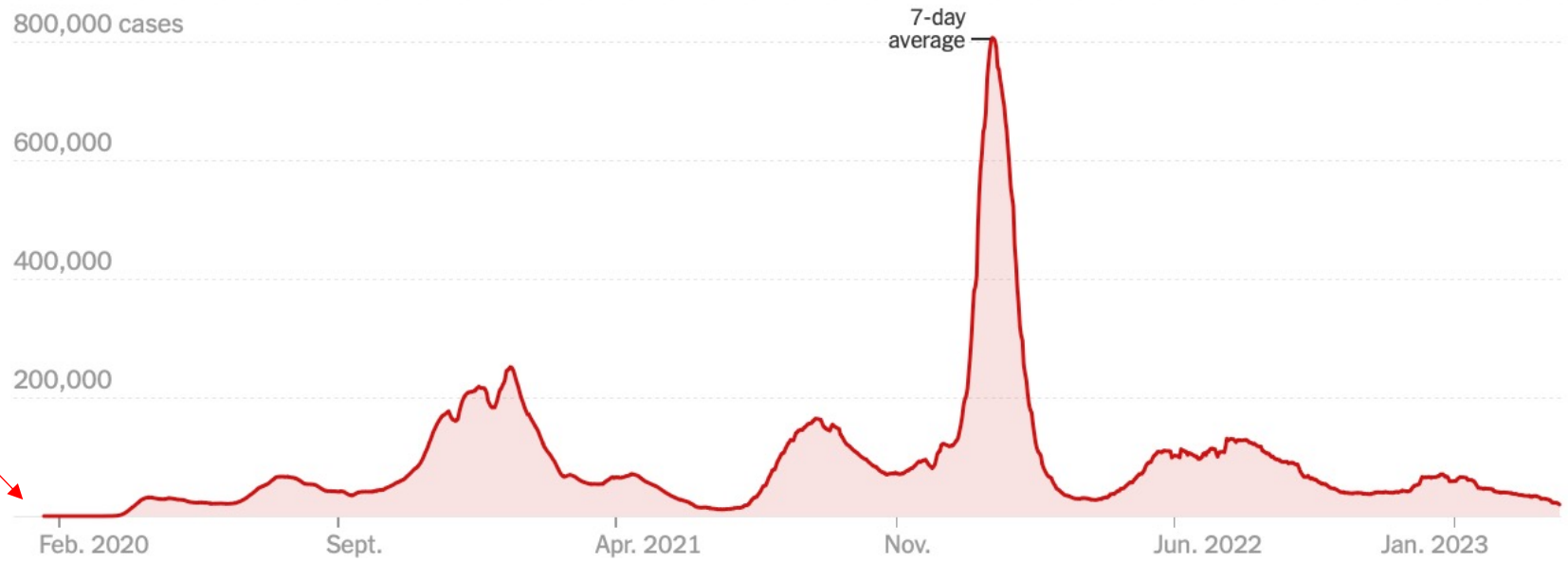
March 24, 2020 at 4:30 p.m. EDT

Washington post – March 2020

Caused significant antibody escape

Can machine learning facilitate *early* and *accurate* predictions of viral escape from antibodies?

A trip back in time to the available data for modeling



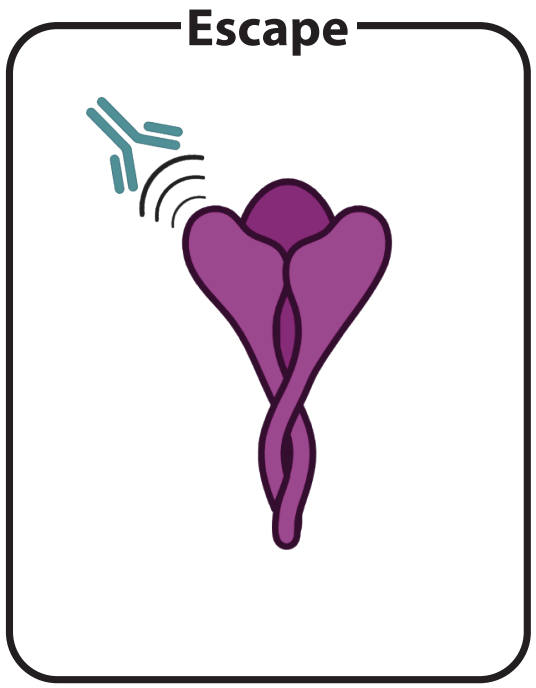
Pandemic onset

>15 million SARS-CoV-2 genomes
>690 Spike-Ab Structures
many experiments

Not using any data after Jan 2020 for model training

nytimes.com

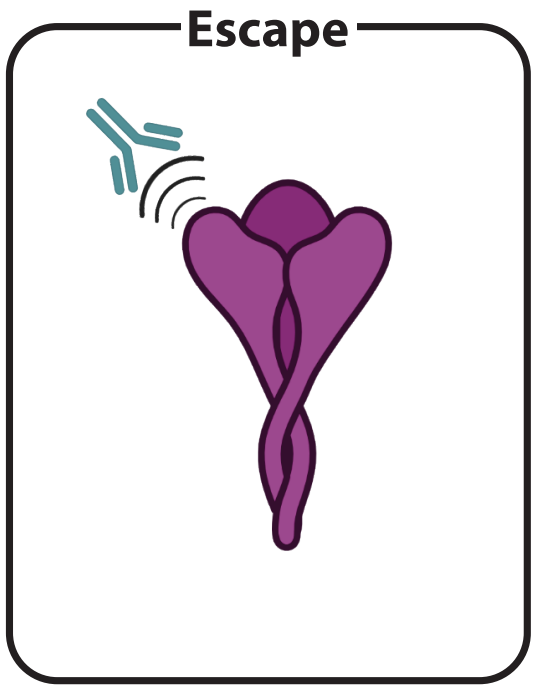
How?



=

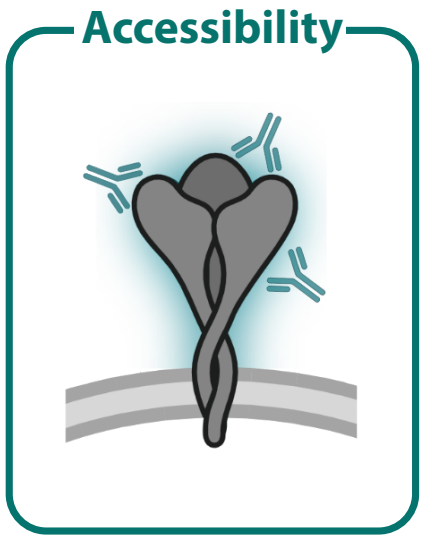
Probability that mutant escapes antibody

How?



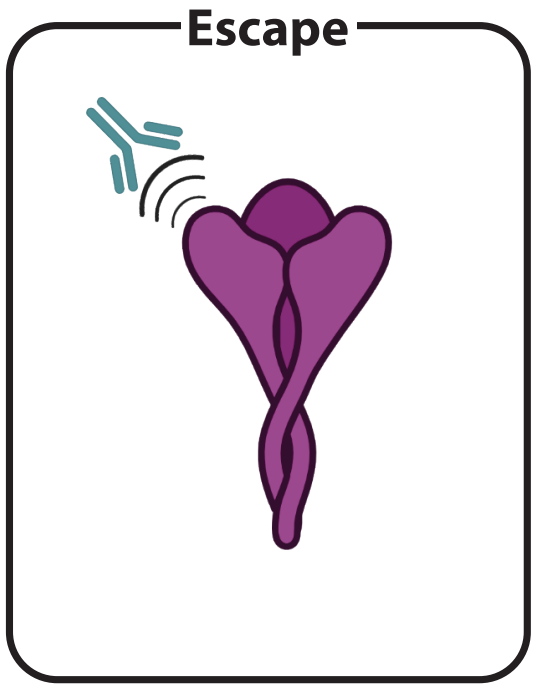
Probability that mutant escapes antibody

=



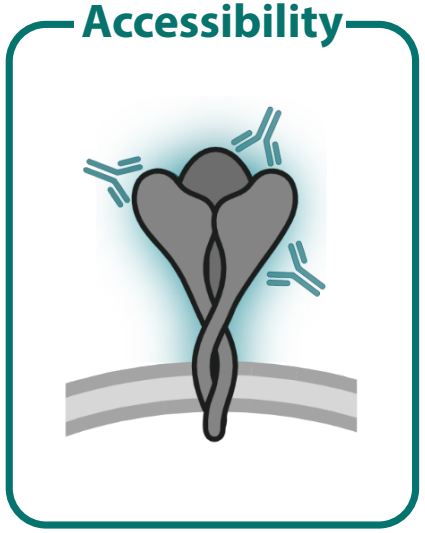
Mutated residue is accessible to antibody

How?



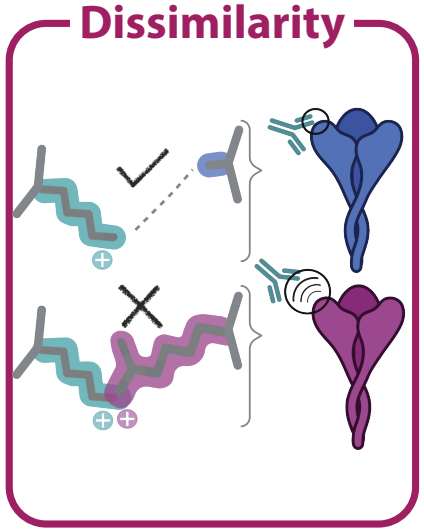
Probability that mutant escapes antibody

=



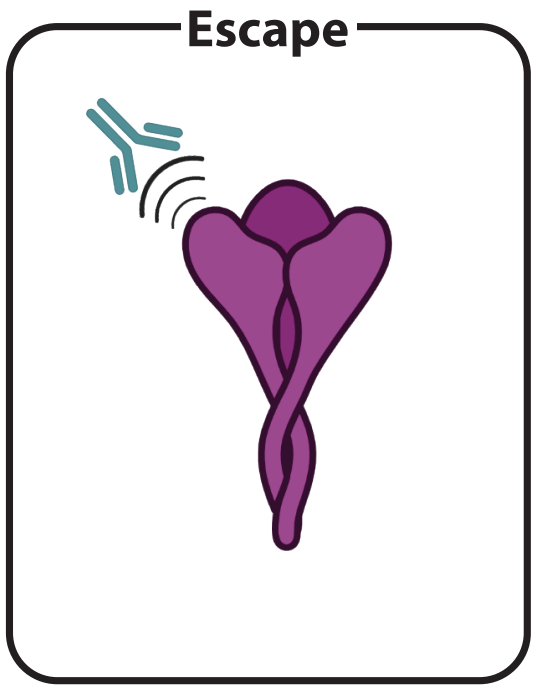
Mutated residue is accessible to antibody

×



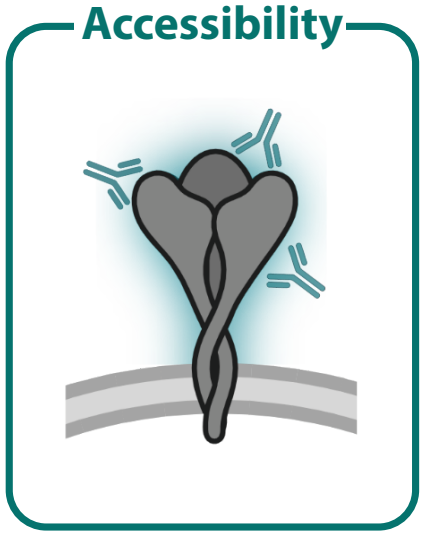
Mutation disrupts antibody binding

How?



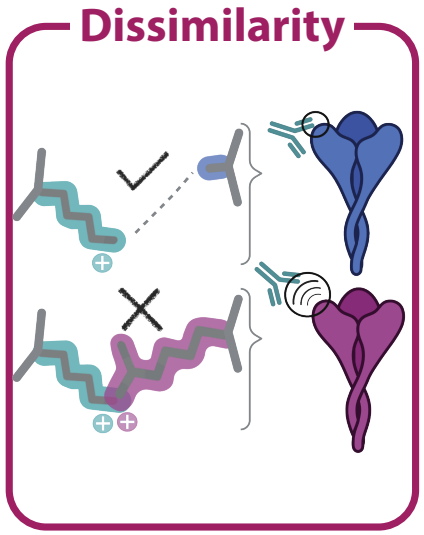
Probability that mutant escapes antibody

=



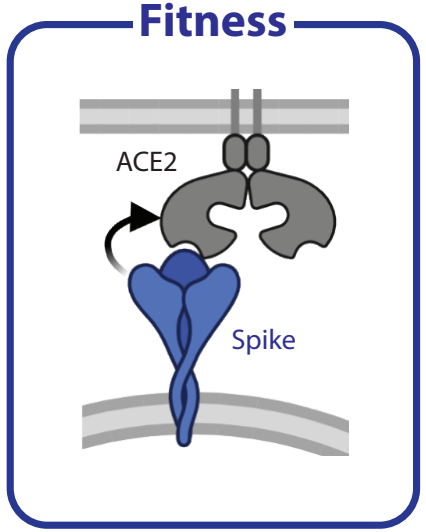
Mutated residue is accessible to antibody

×



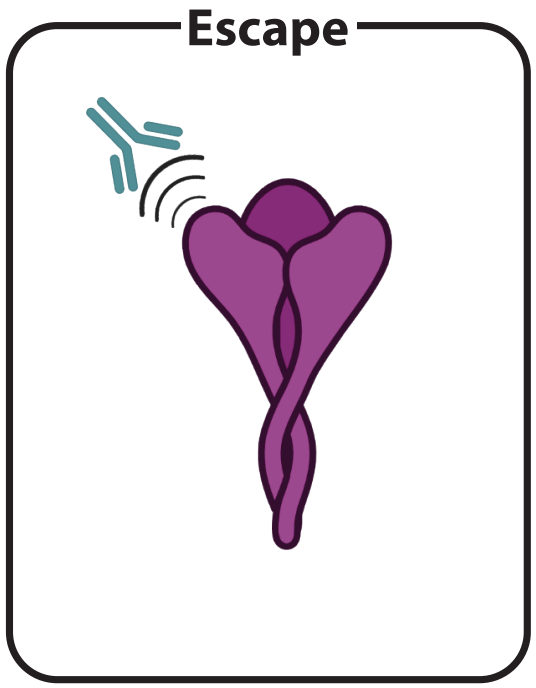
Mutation disrupts antibody binding

×



Mutation maintains viral fitness

How?

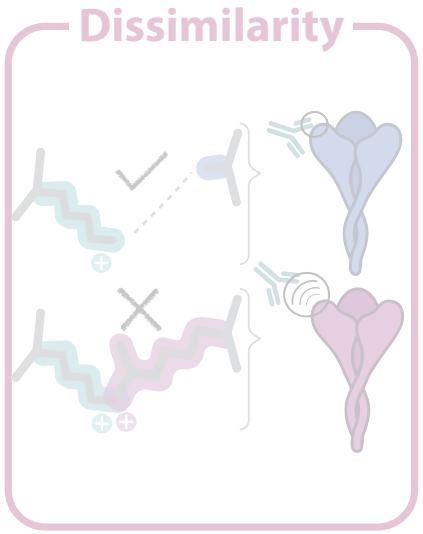


Probability that mutant escapes antibody

=

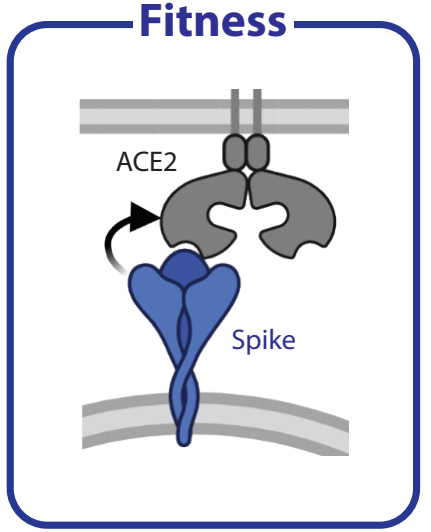
Mutated residue is accessible to antibody

×



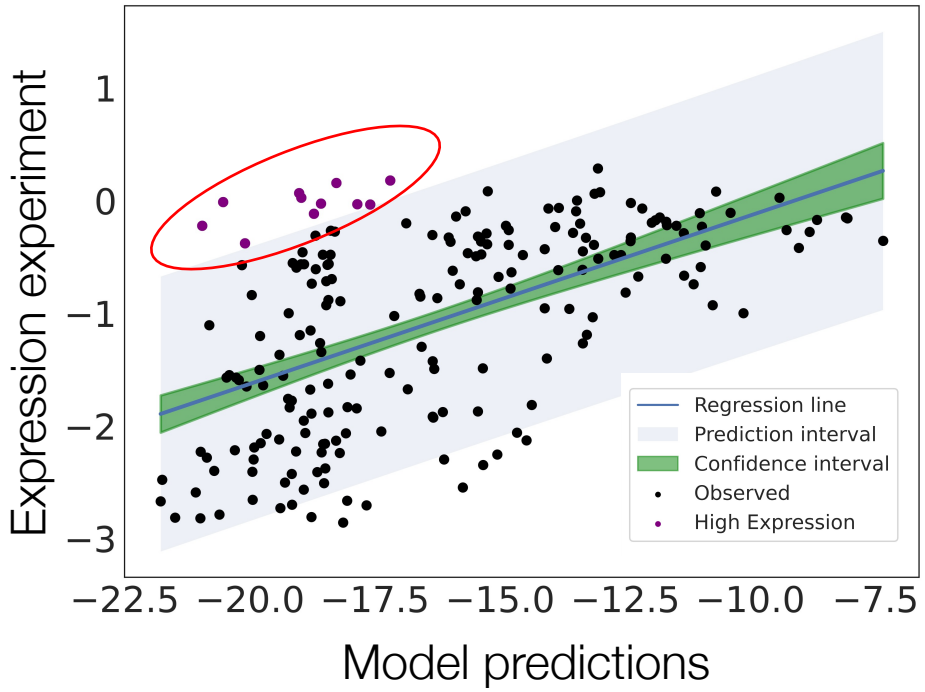
Mutation disrupts antibody binding

×

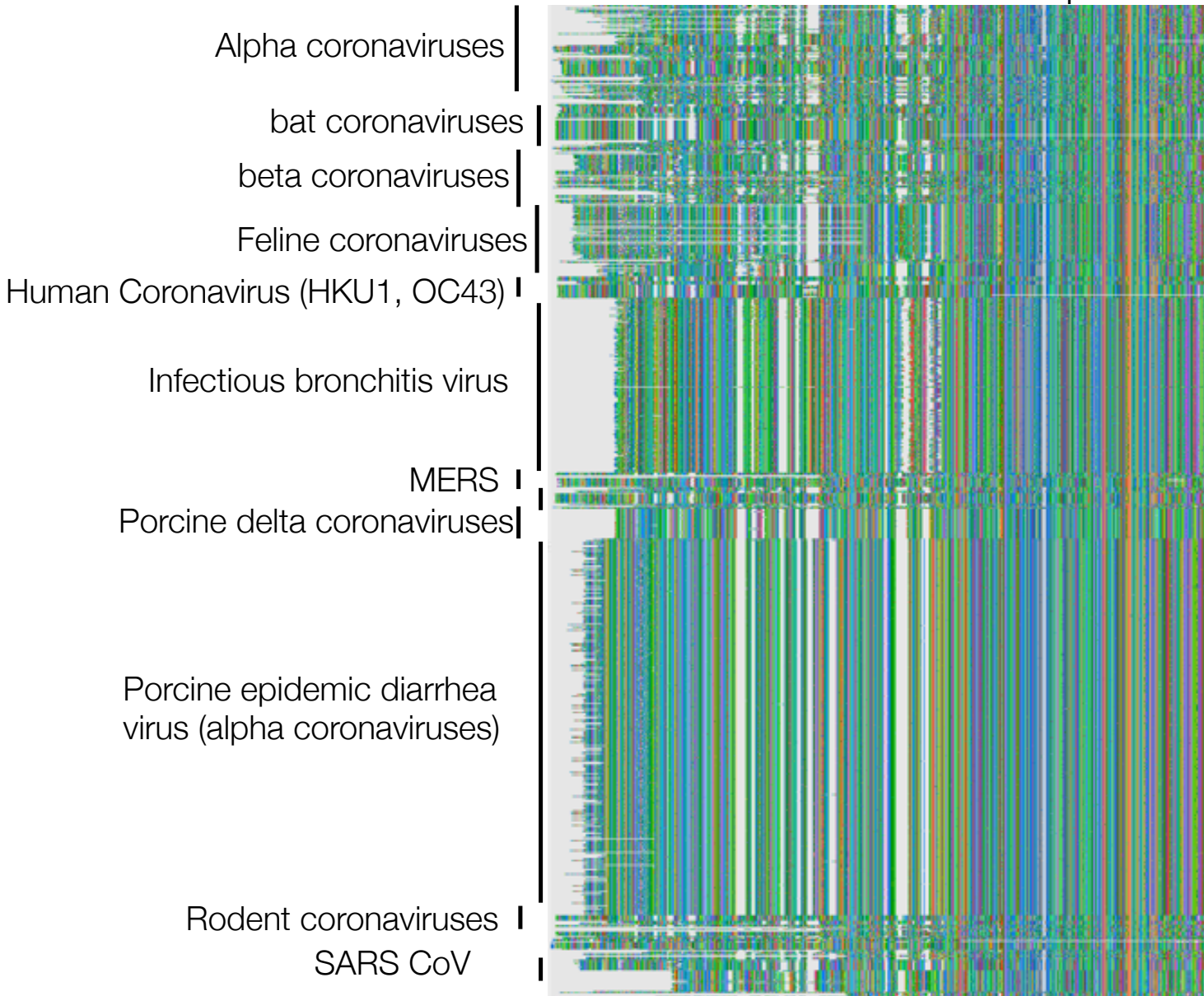


Mutation maintains viral fitness

Deep learning model trained on pre-2020 data predicts SARS-CoV-2 fitness



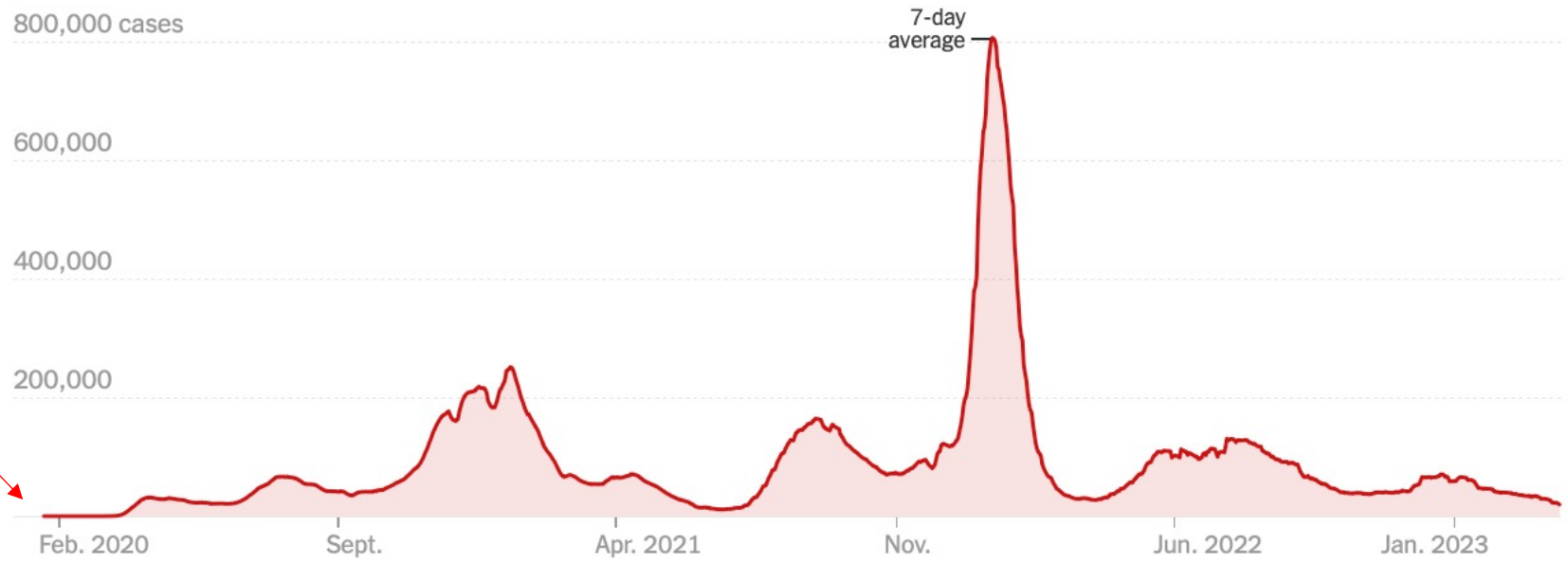
~3k sequences



Starr *et al. Cell,* 2020

Can EVEscape trained
on pre-pandemic data *anticipate pandemic mutations*?

A trip back in time to evaluate predictions



Pandemic onset

>15 million SARS-CoV-2 genomes
>690 Spike-Ab Structures
many experiments

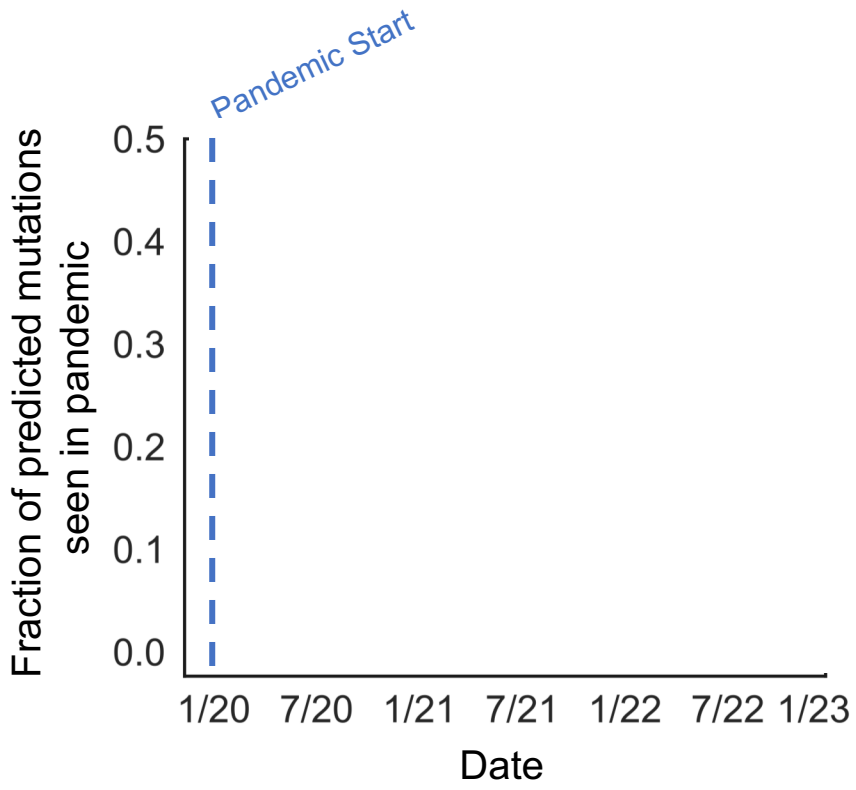
Not using any data after Jan 2020 for model training

Now we can use for evaluation!

nytimes.com

A trip back in time to evaluate predictions

How well do current methods *comprehensively* predict future immune escape?

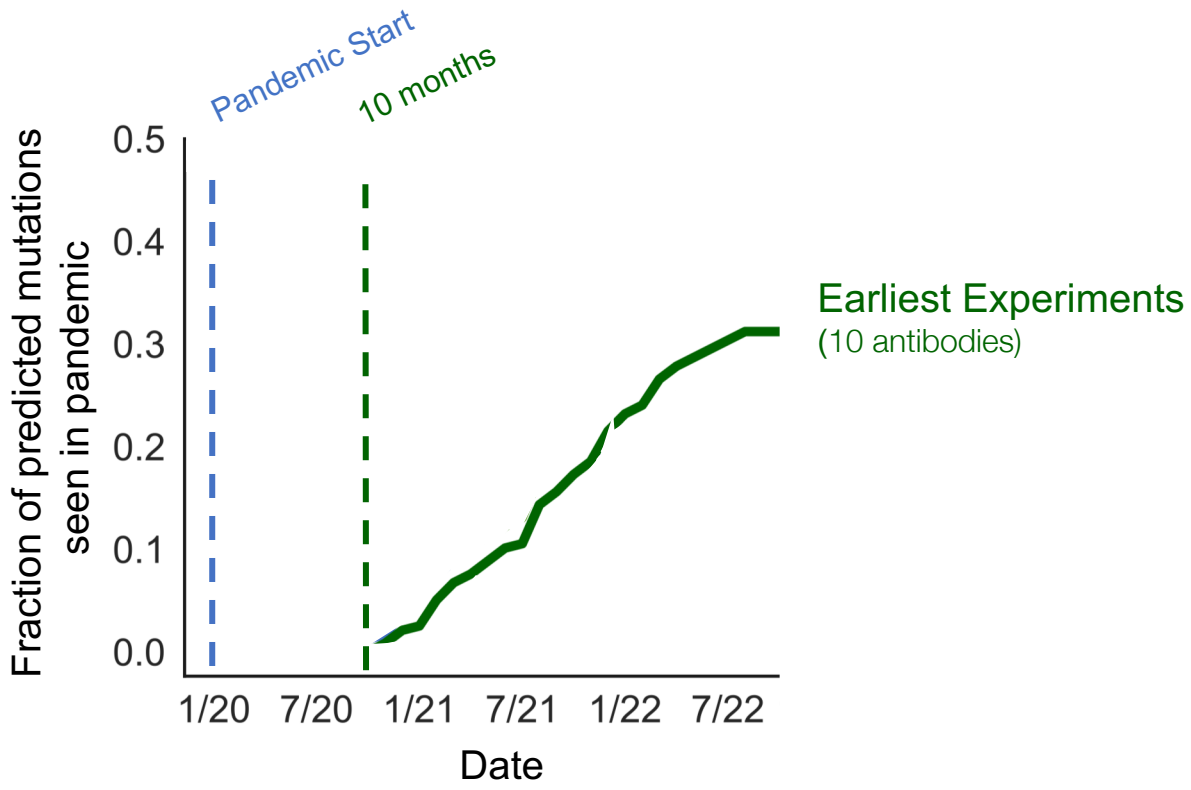


GISAID.org

Thadani*, Gurev*, Notin*… Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

A trip back in time to evaluate predictions

How well do current methods *comprehensively* predict future immune escape?



Neutralization by contemporary antibodies is a weak predictor of future immune escape mutations

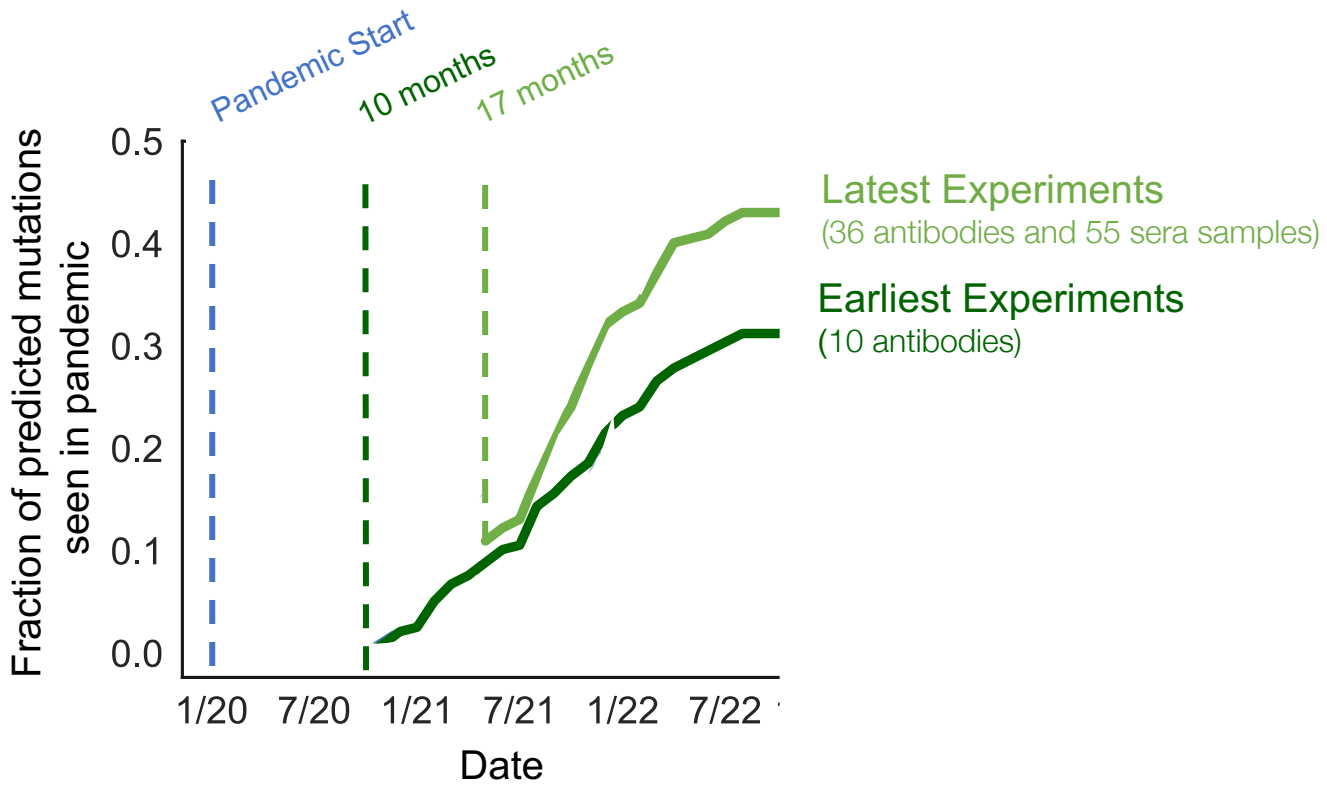
GISAID.org

Thadani*, Gurev*, Notin*… Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

Greaney, Starr, Bloom. An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution,* 2021.

A trip back in time to evaluate predictions

How well do current methods *comprehensively* predict future immune escape?



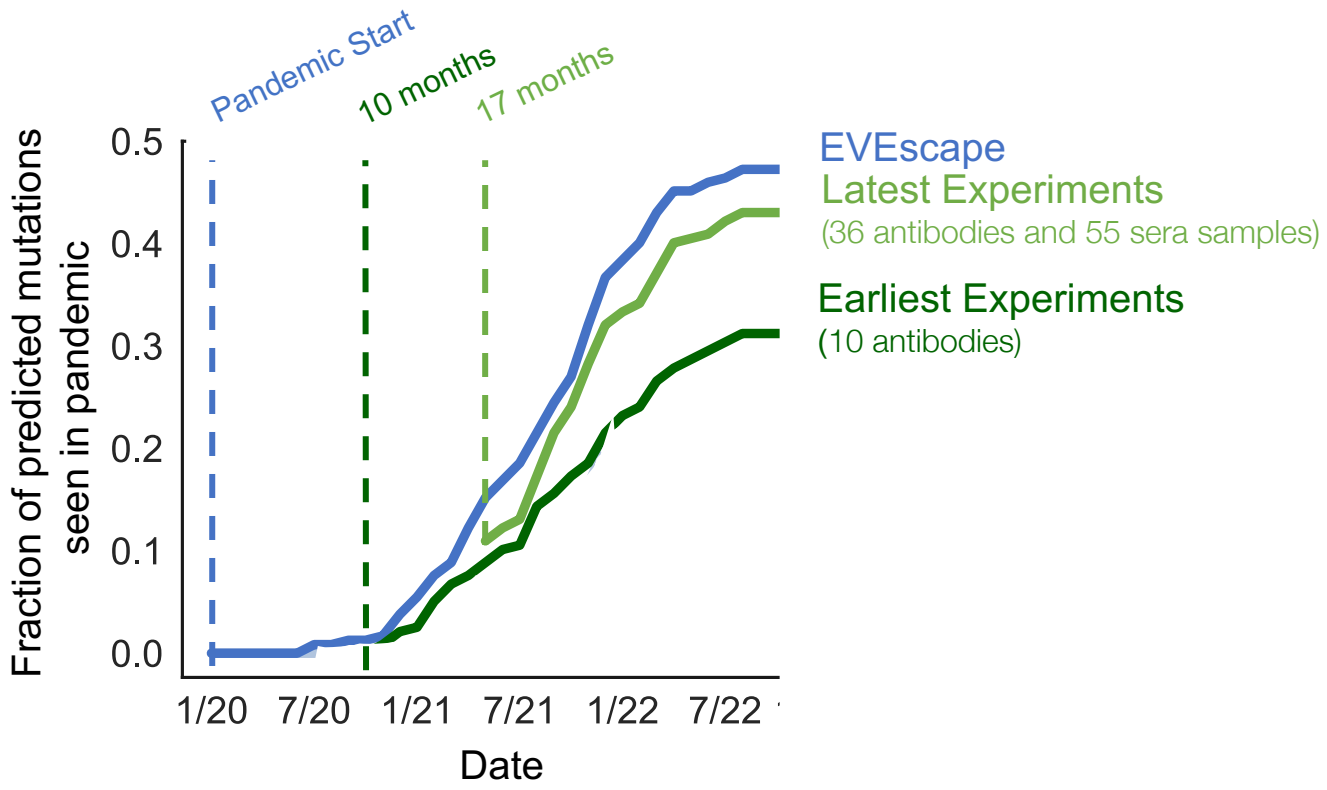
Performance improves with more up-to-date sera

GISAID.org

Thadani*, Gurev*, Notin*... Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

Greaney, Starr, Bloom. An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution,* 2021.

EVEscape available at start of pandemic is a better predictor of viral evolution

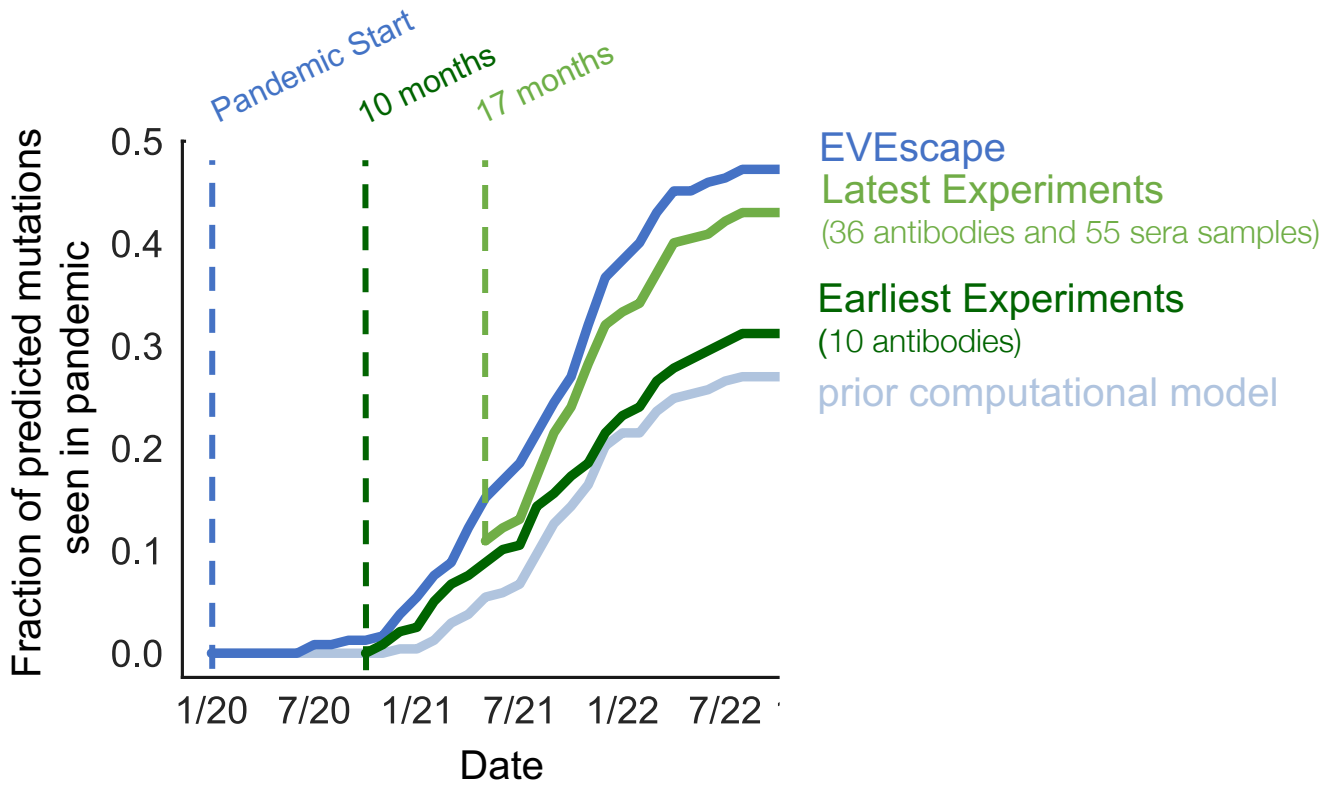


GISAID.org

Thadani*, Gurev*, Notin*... Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

Greaney, Starr, Bloom. An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution,* 2021.

EVEscape available at start of pandemic is a better predictor of viral evolution



GISAID.org

Thadani*, Gurev*, Notin*... Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.
Greaney, Starr, Bloom. An antibody-escape calculator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evolution,* 2021.
Hie, Zhong, Berger, Bryson. Learning the language of viral evolution and escape. *Science*, 2021

Can EVEscape trained
on pre-pandemic data *anticipate pandemic mutations*?

Can EVEscape trained on pre-pandemic data *anticipate pandemic mutations*?

Yes

How can we apply EVEscape in current and future outbreaks?

Early warning of high-escape variants

Evaluate future protection of therapeutics

Vaccine design

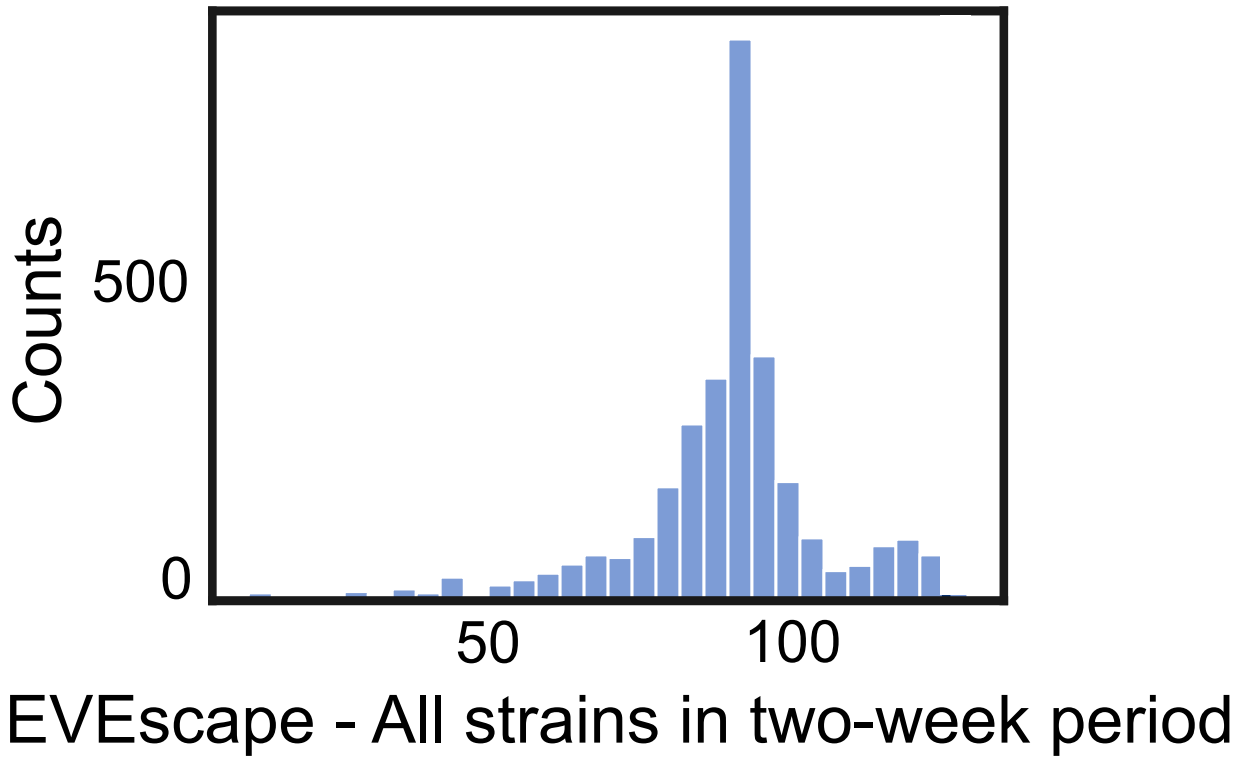
How can we apply EVEscape in current and future outbreaks?

Early warning of high-escape variants

Evaluate future protection of therapeutics

Vaccine design

Thousands of emerging strains are too many to experimentally test

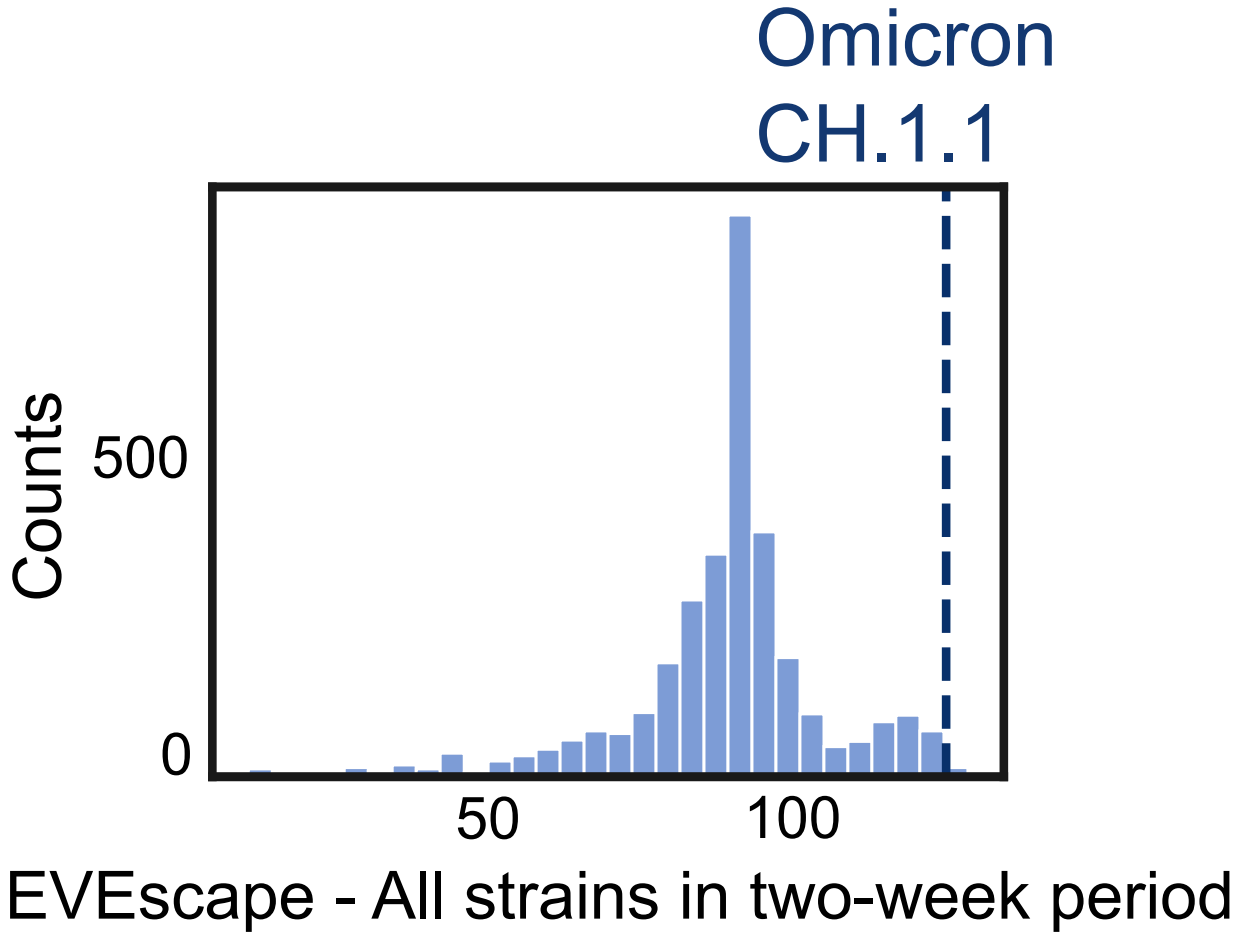


GISAID.org

Thadani*, Gurev*, Notin*... Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

EVEscape identifies Variants of Concern out of thousands of emerging strains

Evescape.org



GISAID.org

Thadani*, Gurev*, Notin*... Marks. Learning from pre-pandemic data to forecast viral escape. *Nature*, 2023.

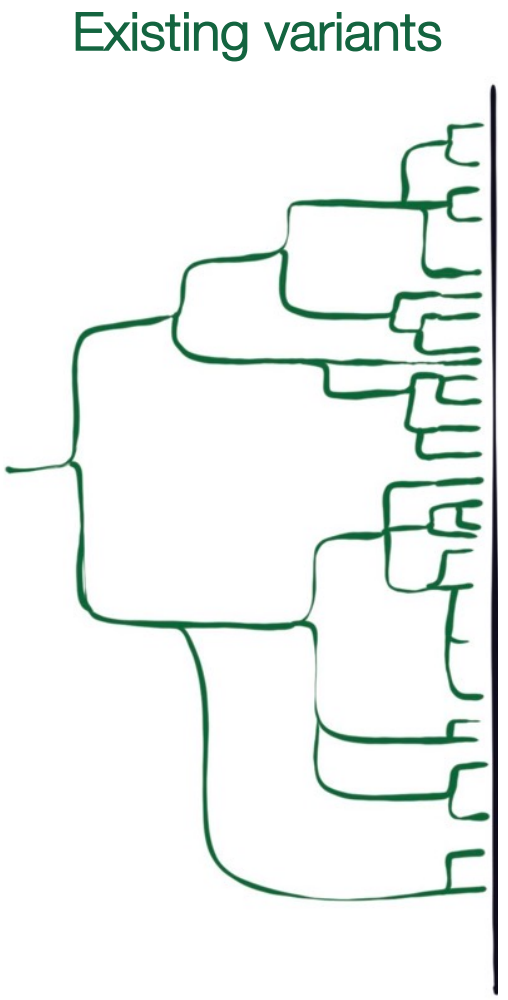
How can we apply EVEscape in current and future outbreaks?

Early warning of high-escape variants

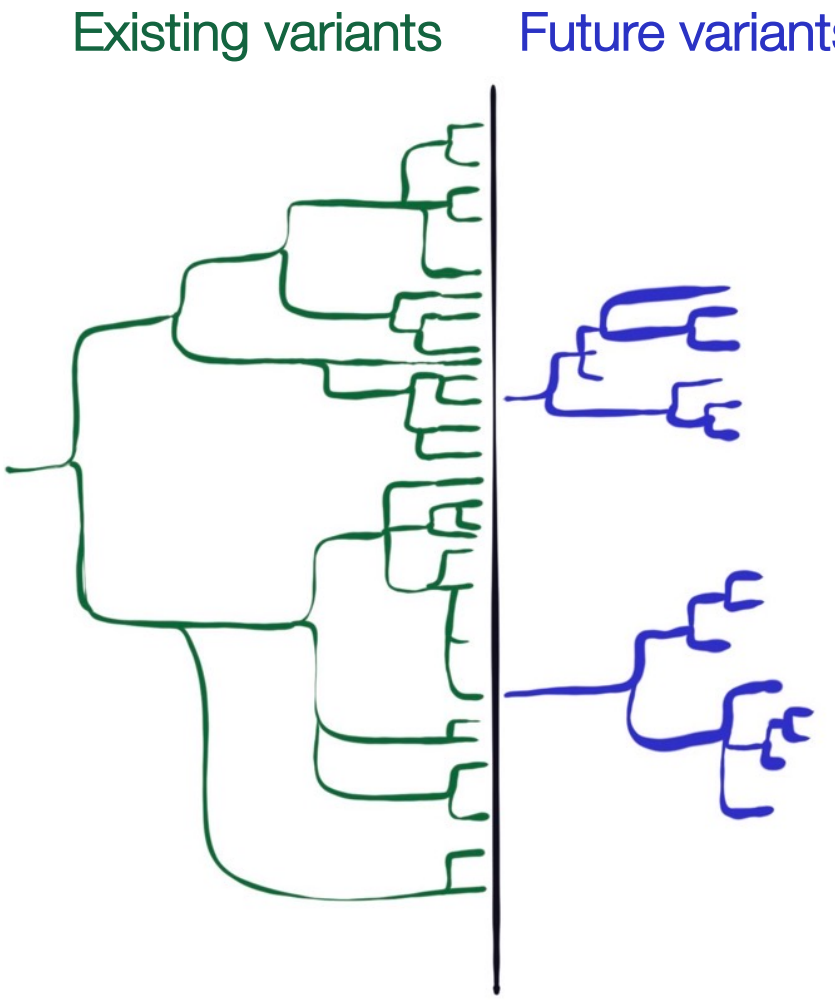
Evaluate future protection of therapeutics

Vaccine design

Assess vaccine efficacy against existing variants

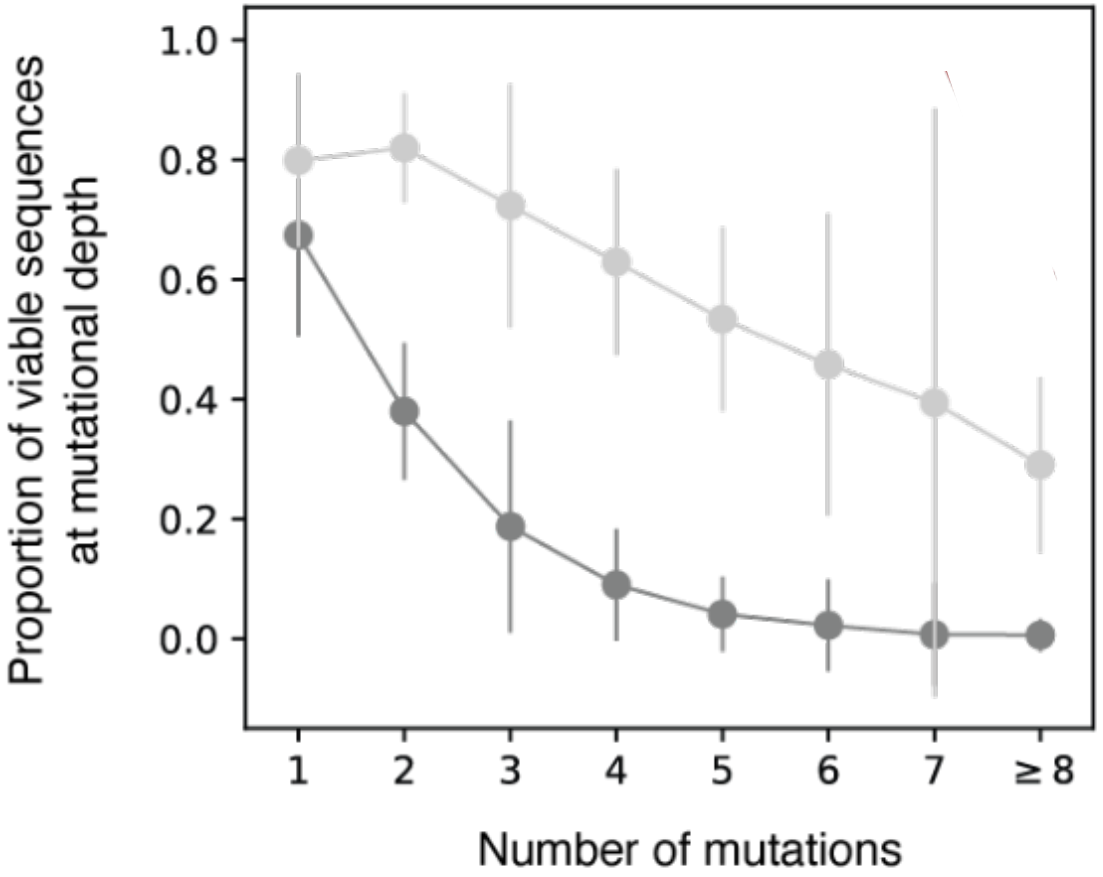


Assess vaccine efficacy against existing variants but also against potential future variants



Why has this been difficult?

Most random multi-mutant proteins are non-viable

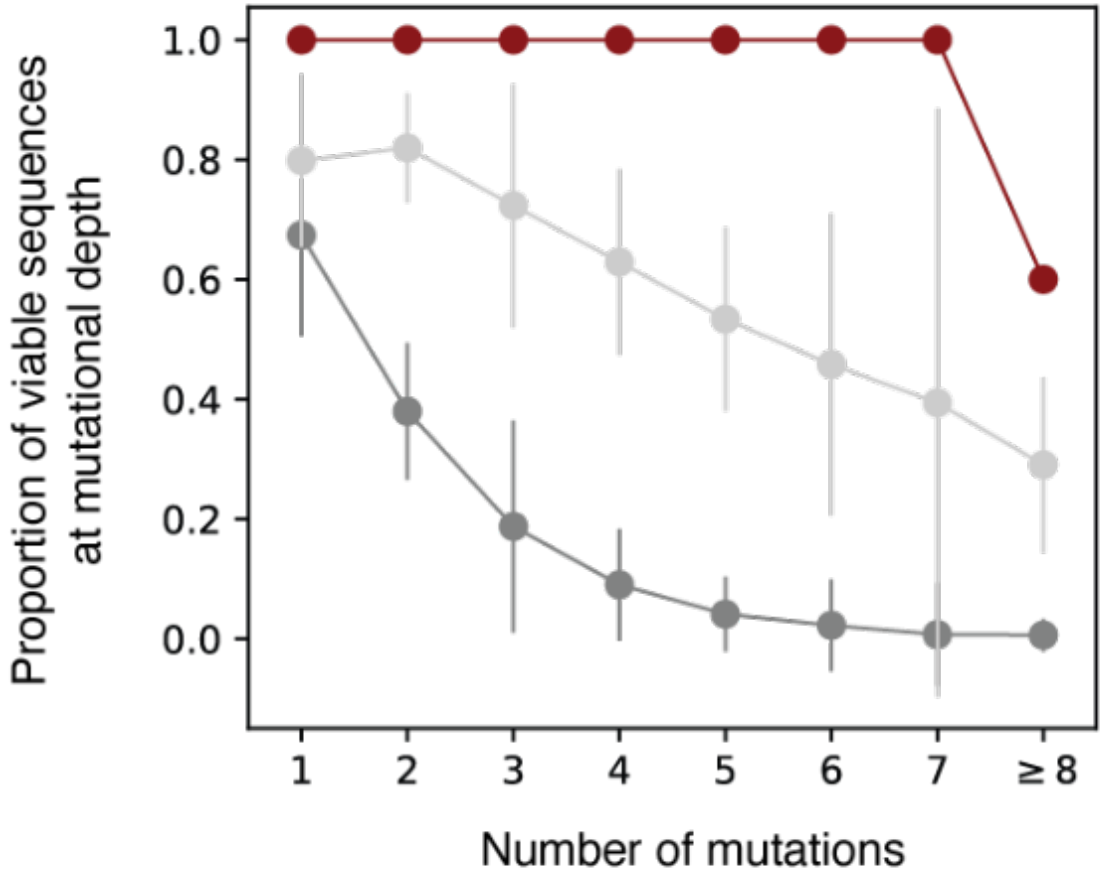


Random mutations

Random pandemic mutations

Starr…Bloom. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell,* 2020.

Designing diverse multi-mutant Spikes using EVEscape



Random mutations

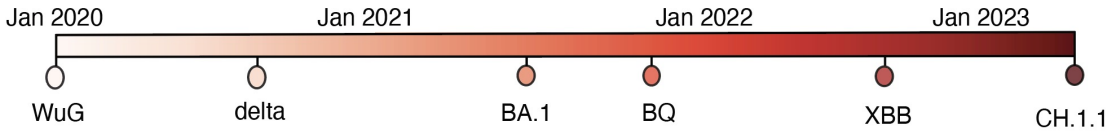
Random pandemic mutations

EVEscape Designs

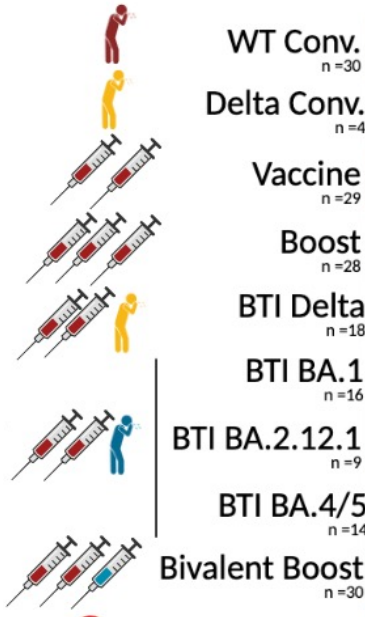
94% of designs are infectious!

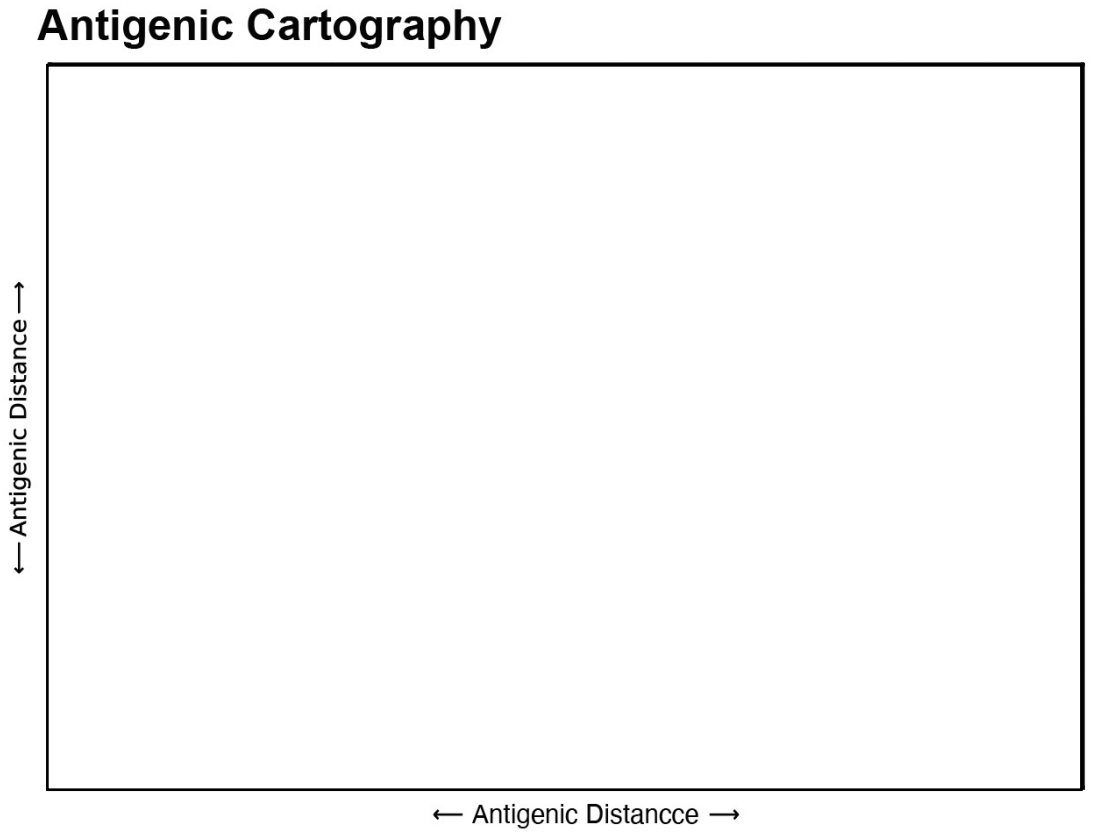
Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.
Starr…Bloom. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell,* 2020.

Designed constructs mimic future evolution



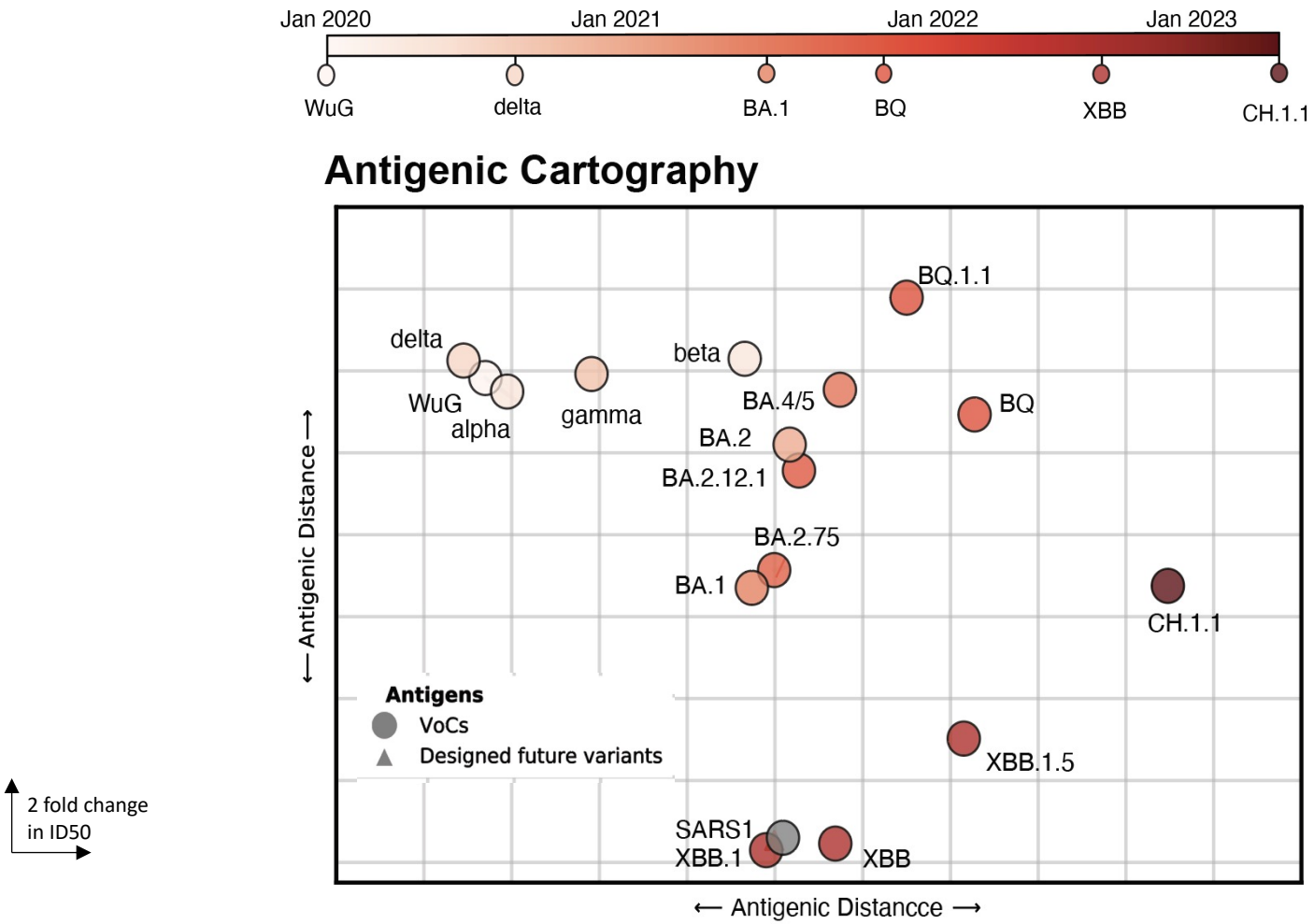
Immunization histories





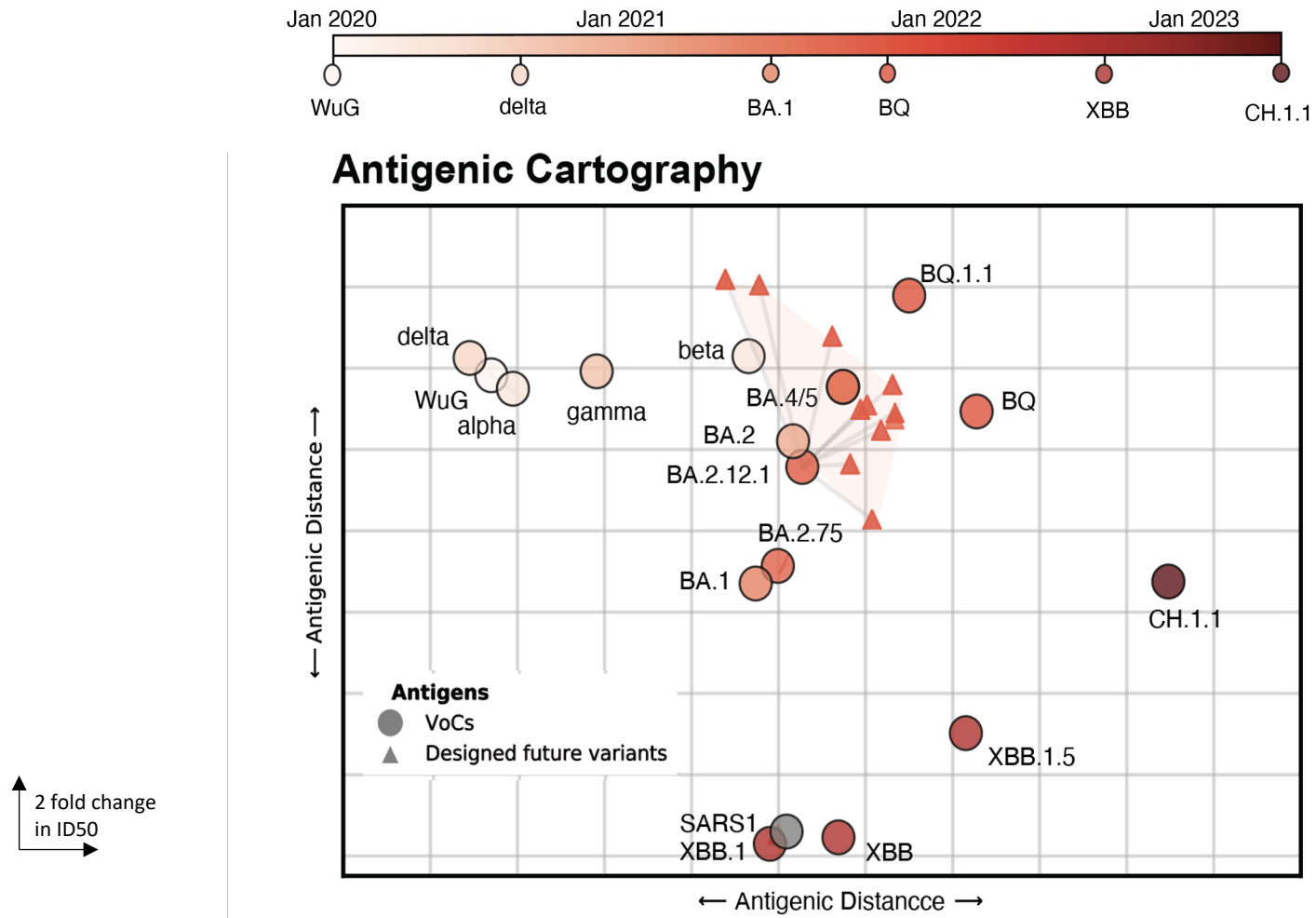
Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.

Designed constructs mimic future evolution



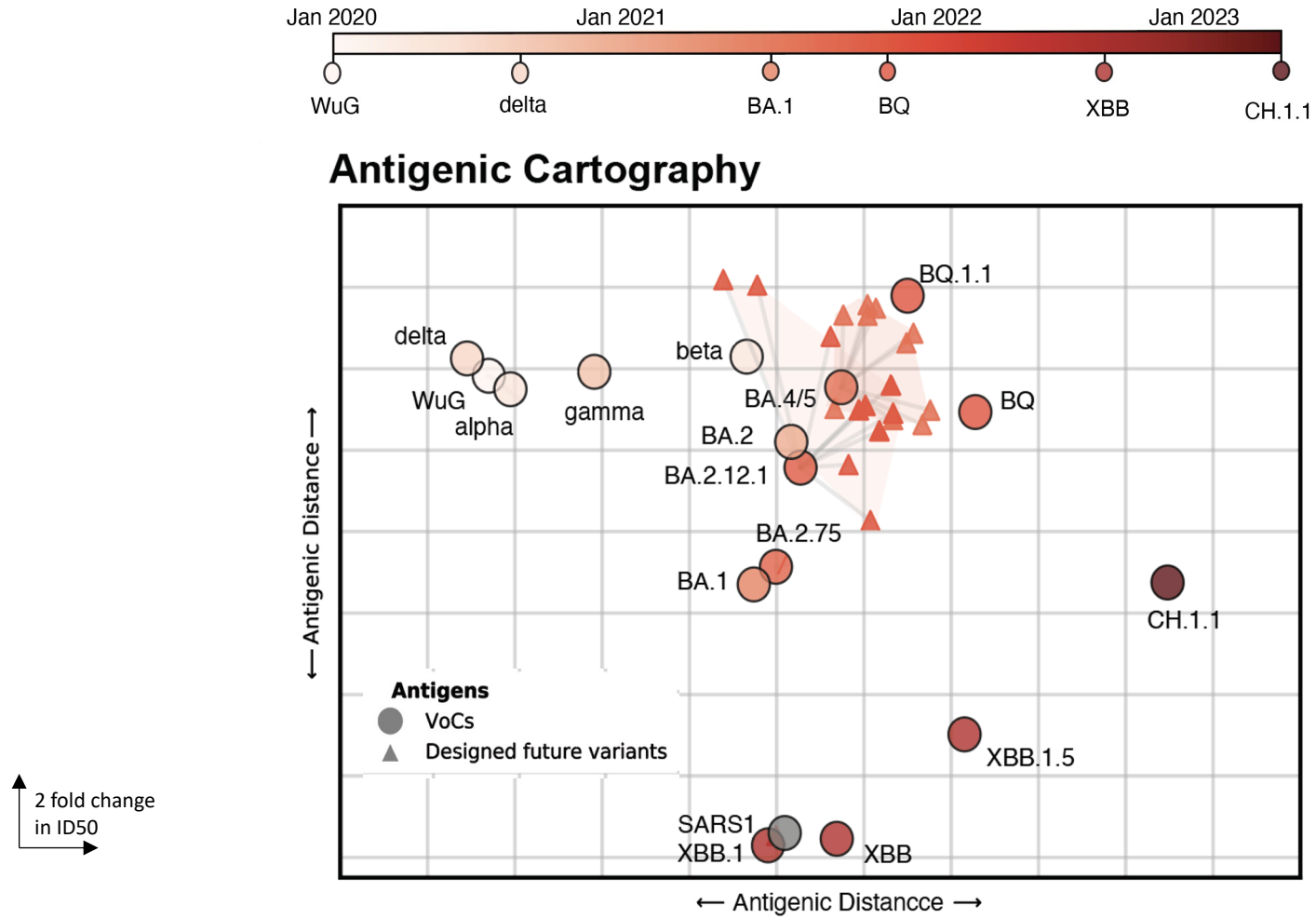
Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.

Designed constructs mimic future evolution



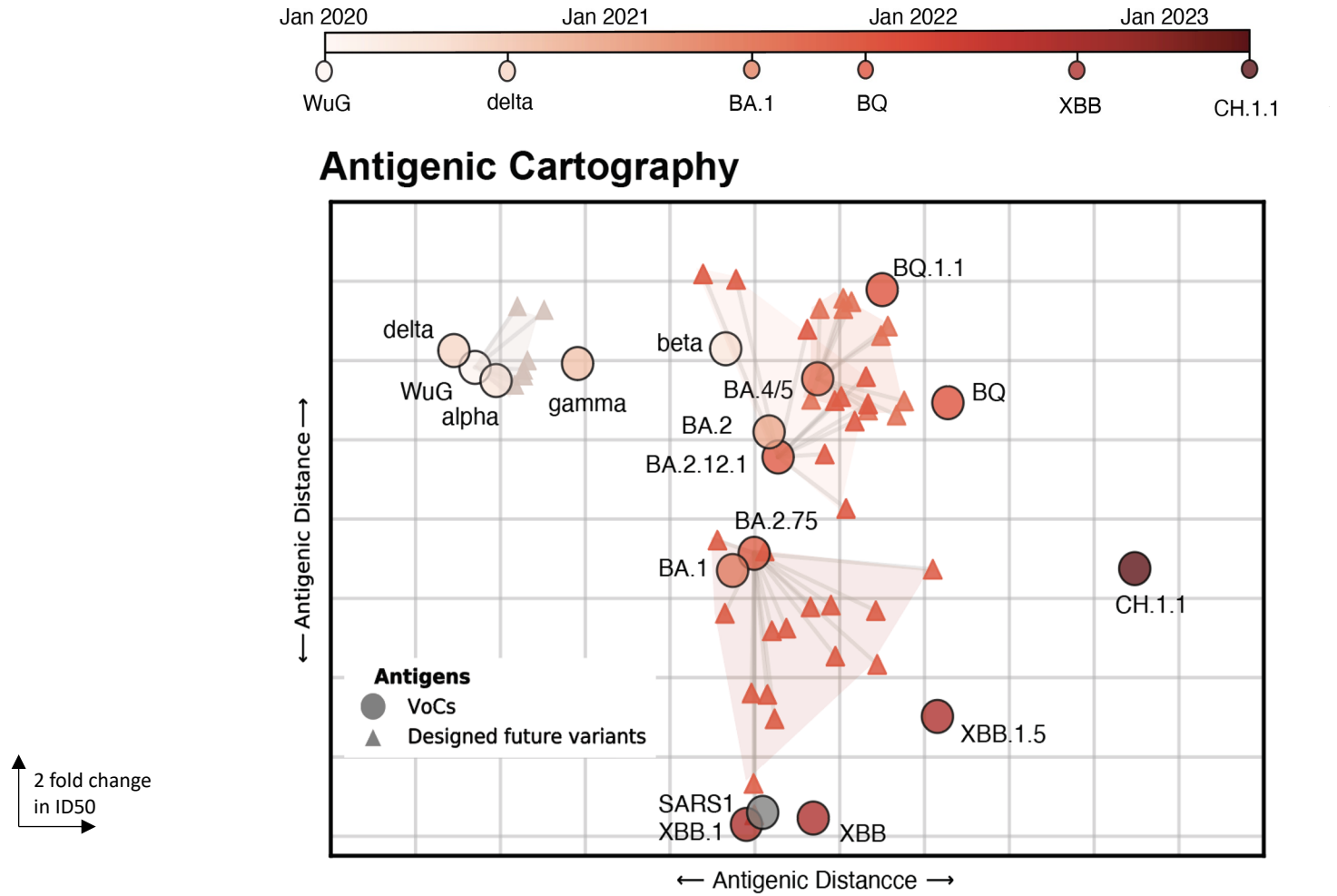
Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.

Designed constructs mimic future evolution



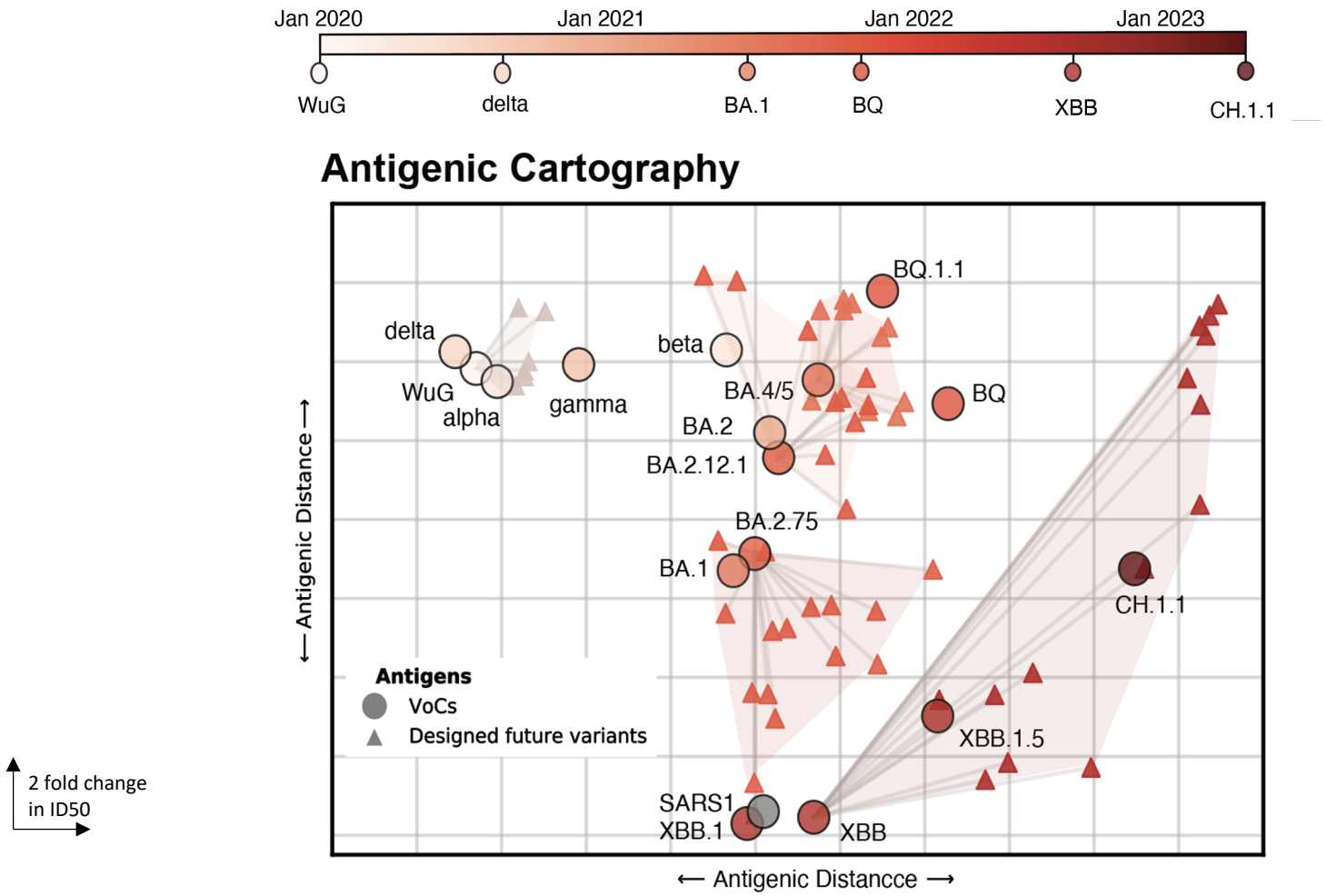
Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.

Designed constructs mimic future evolution



Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Biorxiv*, 2023.

Designed constructs mimic future evolution



Youssef…Lemieux, Luban, Seaman, Marks. Deep generative models predict SARS-CoV-2 Spike infectivity and foreshadow neutralizing antibody escape. *Bioxiv*, 2023.

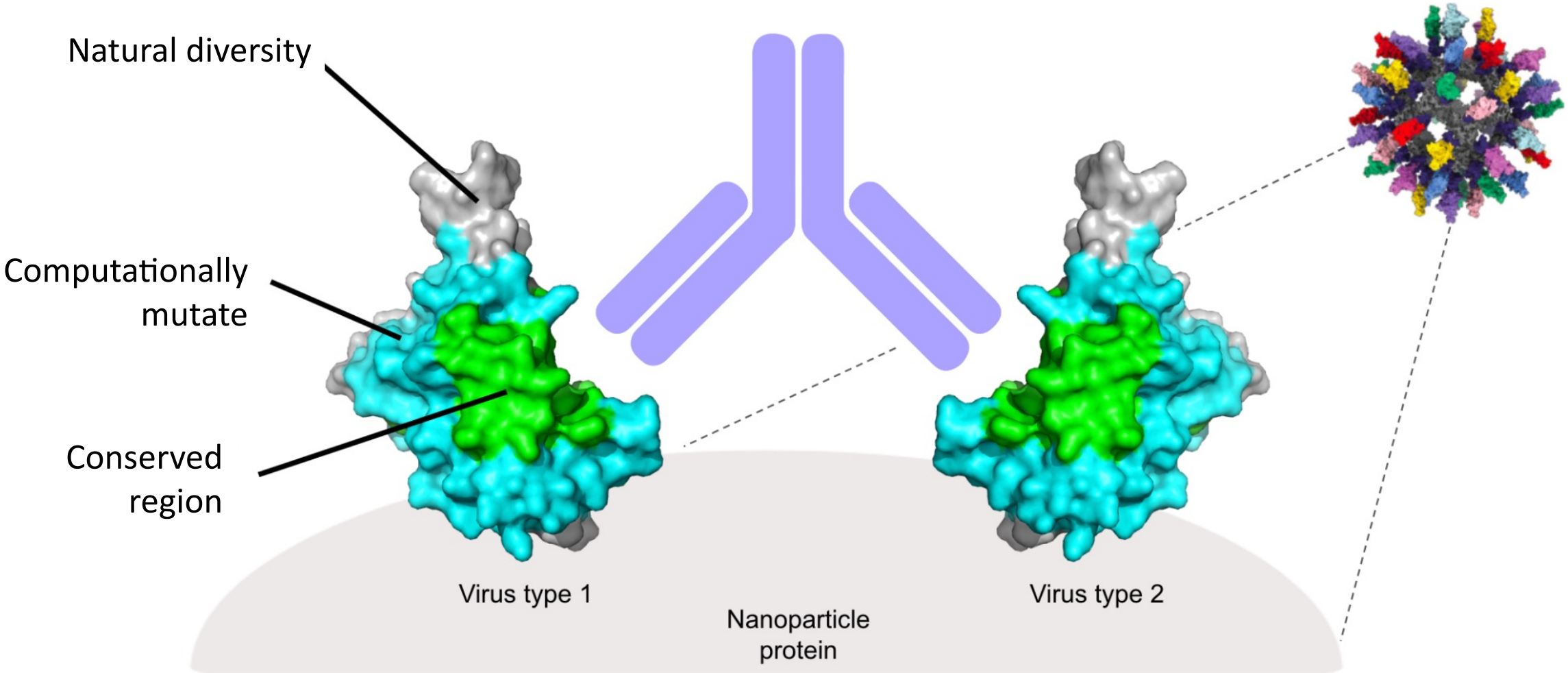
How can we apply EVEscape in current and future outbreaks?

Early warning of high-escape variants

Evaluate future protection of therapeutics

Vaccine design

Computationally designed vaccines to focus antibodies to regions unlikely to mutate in the future



Gurev,* Youssef*…Marks. Future-proof vaccine design with a generative model of antibody cross-reactivity. ICLR GEMbio, 2023.