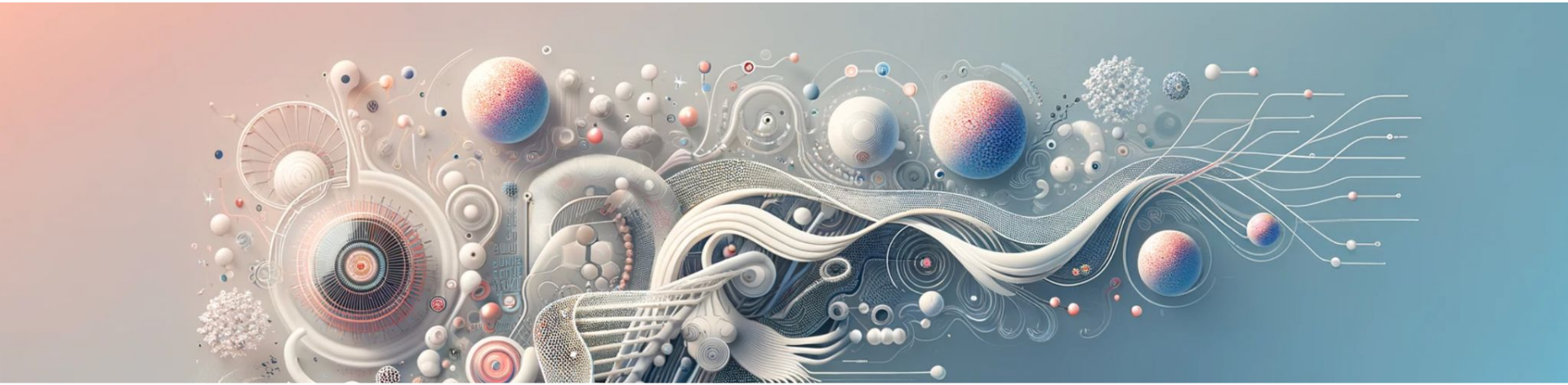
CSE7850/CX4803 Machine Learning in Computational Biology

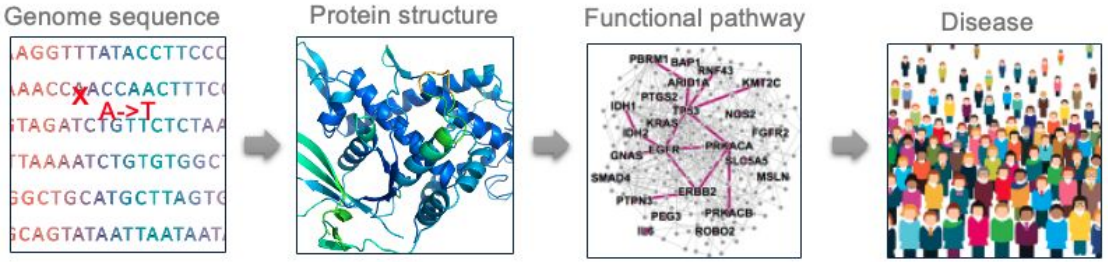


Lecture 17: Variant Effect Prediction

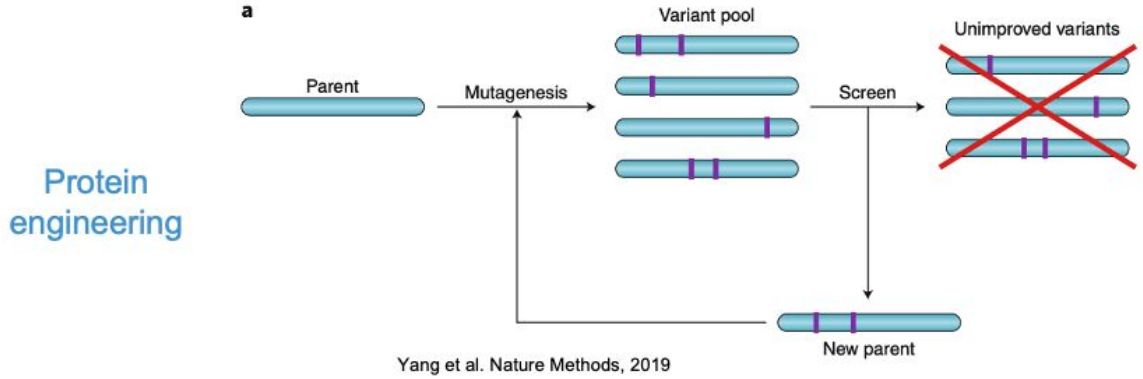
Yunan Luo

Understanding the effect of mutations/variants

- Identify “bad” mutations (pathogenic variants in human disease)

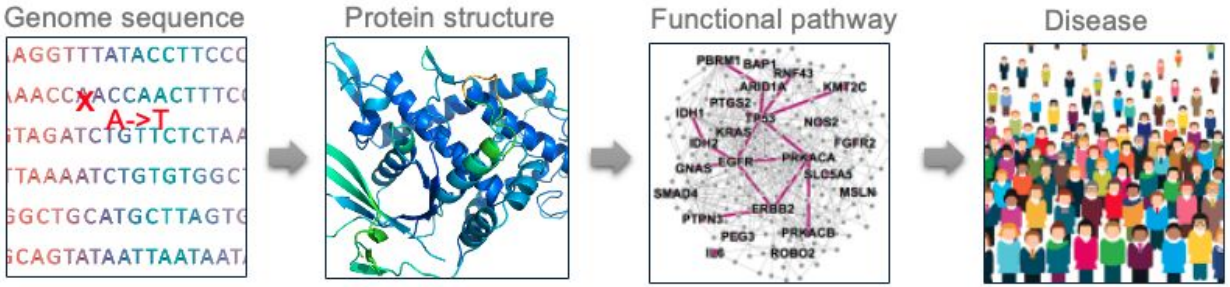


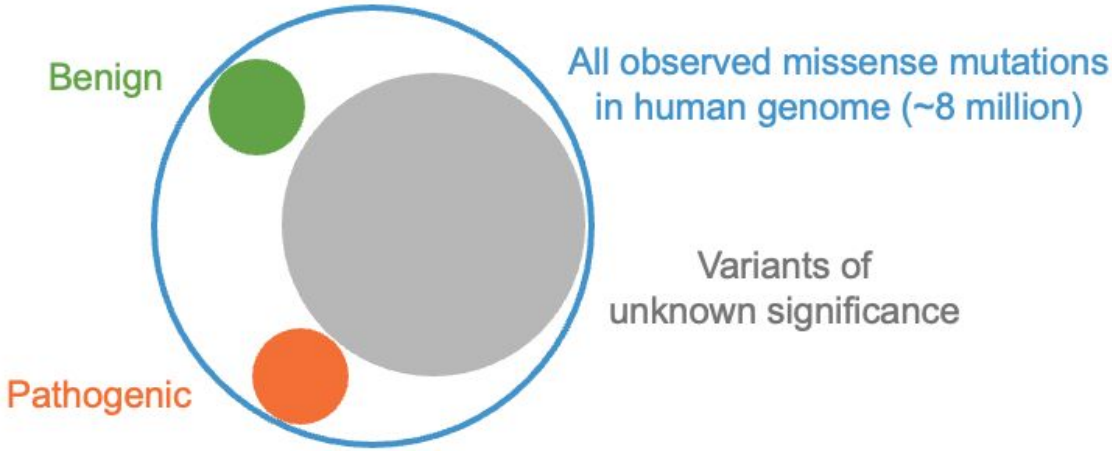
- Identify “good” mutations (function-enhancing variants in protein engineering)



Yang et al. Nature Methods, 2019

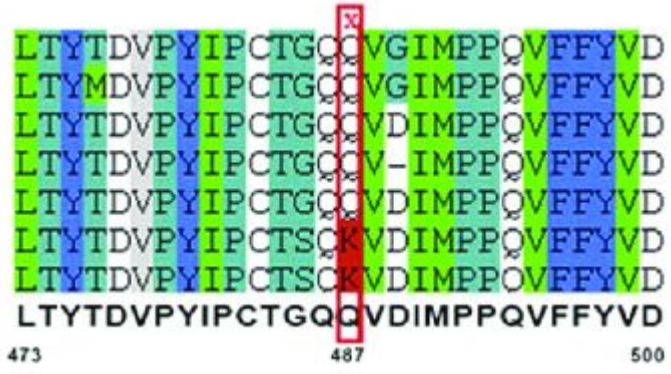
Pathogenicity of disease variants





How to predict pathogenicity without labeled data?

- Evolutionary information revealed by sequence alignment:



- Any limitation of this method?

How to capture pairwise dependencies?

Multiple sequence alignment

$$P(\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_i e_i(x_i) + \sum_{i\neq j} e_{ij}(x_i, x_j)\right)$$

Single potentials

Local preference

Pairwise potentials

Co-evolution strength

Markov random field
Ising (Potts) model
Undirected graphical model

How to capture more than 2-order dependencies

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_i e_i(x_i) + \sum_{i \neq j} e_{ij}(x_{ij})\right)$$

- Co-evolution based methods only consider second order interaction
  - Can we extend to model higher-order terms?

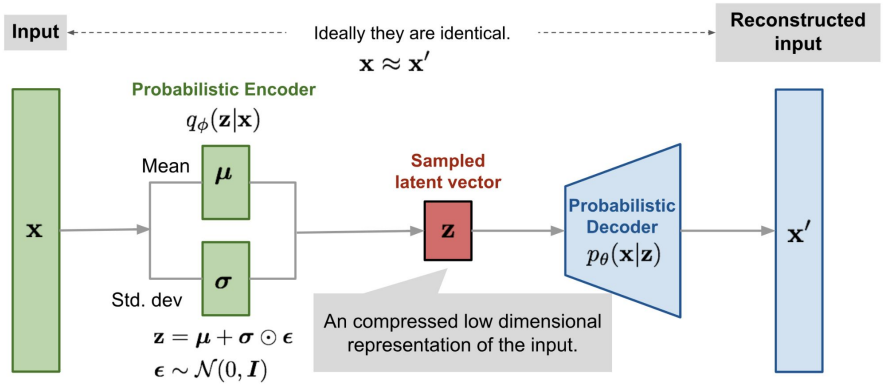
$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\left(\sum_i e_i(x_i) + \sum_{i \neq j} e_{ij}(x_{ij}) + \sum_{i \neq j \neq k} e_{ijk}(x_{ijk}) + \cdots\right)$$

- Computationally hard
  - An exponential number of terms

- Latent-variable ML models capture higher-order interactions *implicitly*

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

Pseudocode of VAE




```python
class variational_autoencoder(nn.Module):
  def __init__(self, ...):
    super(variational_autoencoder,self).__init__()
    self.encoder = Encoder(...)
    self.decoder = Decoder(...)

  def forward(self, x):
    mu, sigma = self.encoder(x)
    z = self.sample(mu, sigma)
    x_prime = self.decoder(z)
    return x
```

- Objective function: $\max \mathbb{E}_q [\log p(\mathbf{x}|\mathbf{z})] - \mathrm{D}_{\mathrm{KL}}(q\|p)$
  - **Reconstruction consistency**: $\mathbb{E}_q[\log p(\mathbf{x}|\mathbf{z})] = -\frac{1}{2\sigma^2}\mathbb{E}_q[\|\mathbf{x} - G_\theta(\mathbf{z})\|^2] + \text{const}$
  - **KL divergence between *p*(*z*) and *q*(*z*)**: typically, p(z) = N(0, 1); The KL term encourages *q*(*z*) to be close to the standard normal distribution N(0, 1)

Paper #2

Article

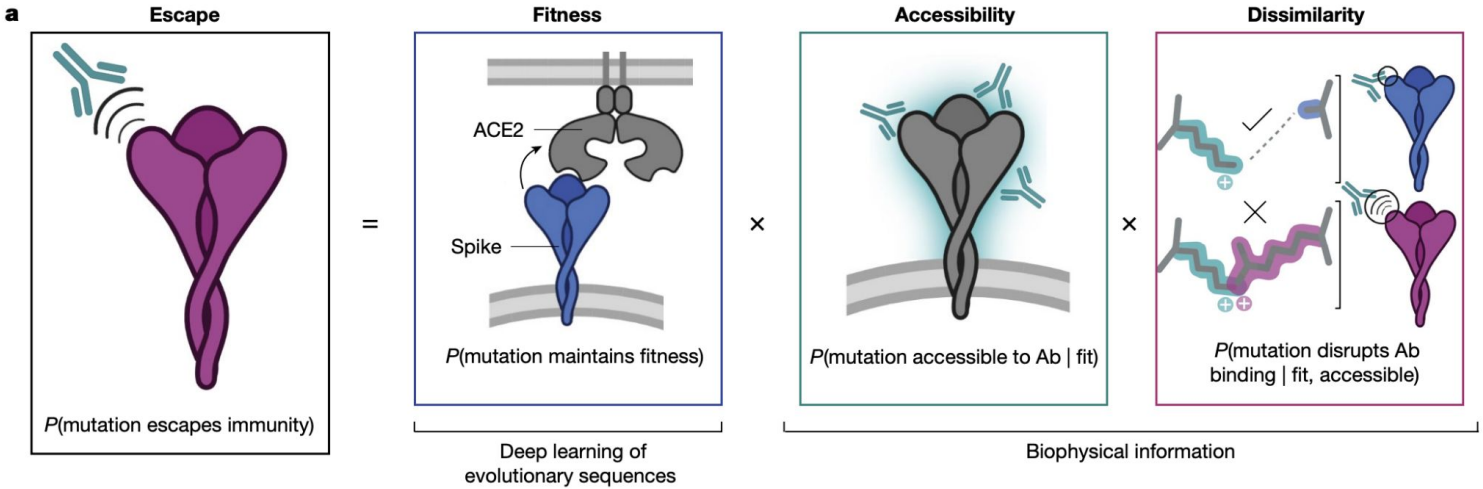
Learning from prepandemic data to forecast viral escape

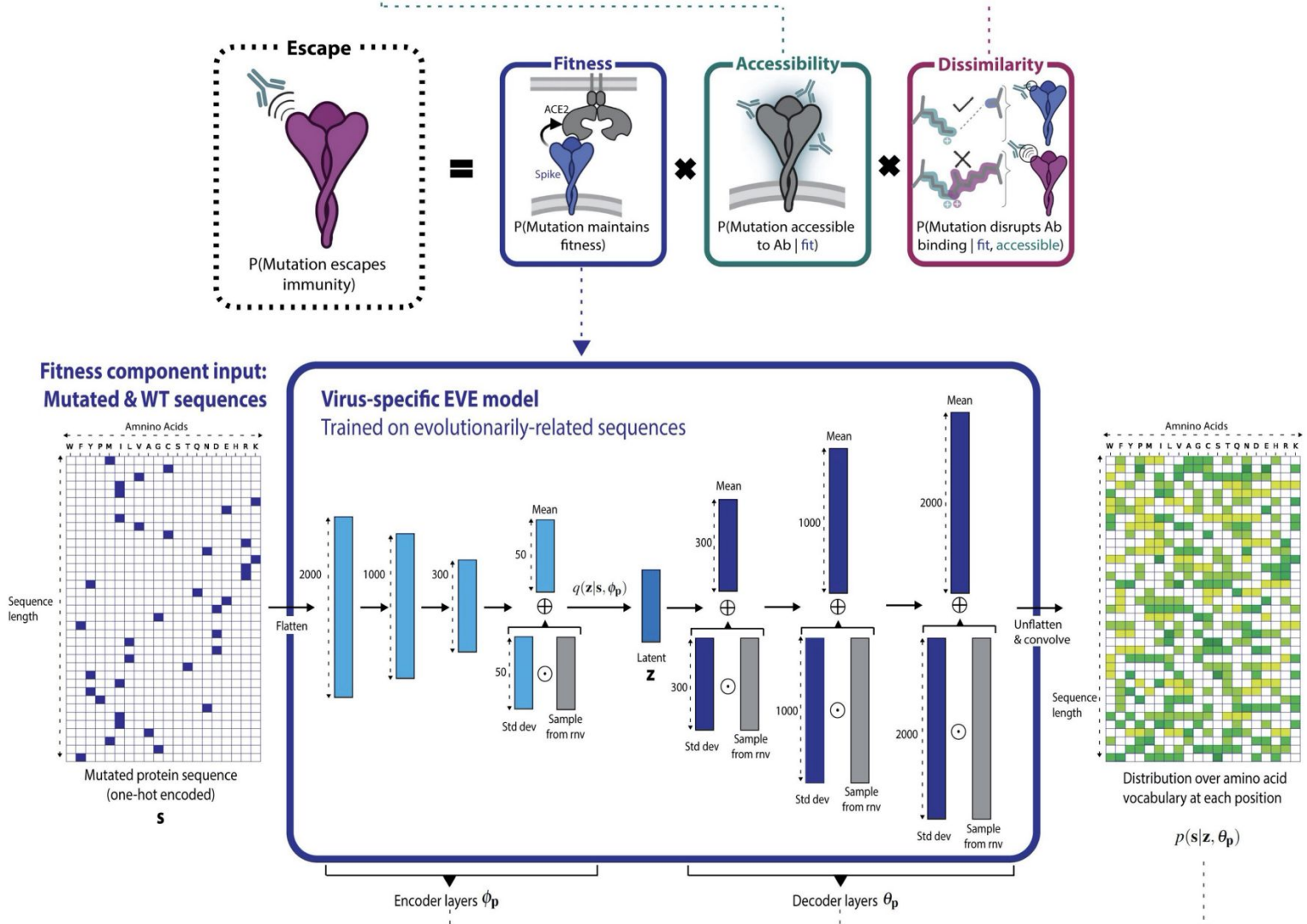
https://doi.org/10.1038/s41586-023-06617-0

Received: 20 July 2022

Accepted: 6 September 2023

Nicole N. Thadani[1,6], Sarah Gurev[1,2,6], Pascal Notin[3,6], Noor Youssef[1], Nathan J. Rollins[1,5], Daniel Ritter[1], Chris Sander[1,4], Yarin Gal[3] & Debora S. Marks[1,4]

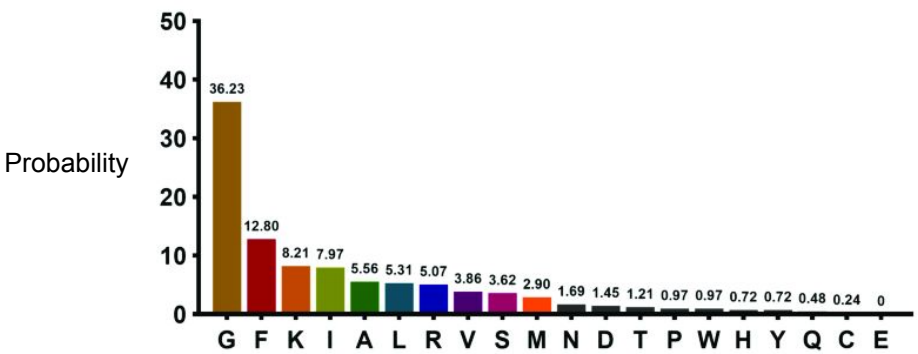



Escape
P(Mutation escapes immunity)
=
Fitness
ACE2
Spike
P(Mutation maintains fitness)
×
Accessibility
P(Mutation accessible to Ab | fit)
×
Dissimilarity
P(Mutation disrupts Ab binding | fit, accessible)
Fitness component input: Mutated & WT sequences
Amino Acids
W F Y P M I L V A G C S T Q N D E H R K
Sequence length
Mutated protein sequence (one-hot encoded)
s
Virus-specific EVE model
Trained on evolutionarily-related sequences
Flatten
2000
1000
300
Mean
50
Std dev
Sample from rnv
$q(\mathbf{z}|\mathbf{s},\phi_\mathbf{p})$
Latent
z
Unflatten & convolve
Encoder layers $\phi_\mathbf{p}$
Decoder layers $\theta_\mathbf{p}$
Distribution over amino acid vocabulary at each position
$p(\mathbf{s}|\mathbf{z},\theta_\mathbf{p})$

Another way to predict mutation effect: language model

Masked language models

$$p(x) = \prod_{i=1}^{L} p(x_i | x_1 \ldots x_{i-1}, x_{i+1} \ldots x_L)$$

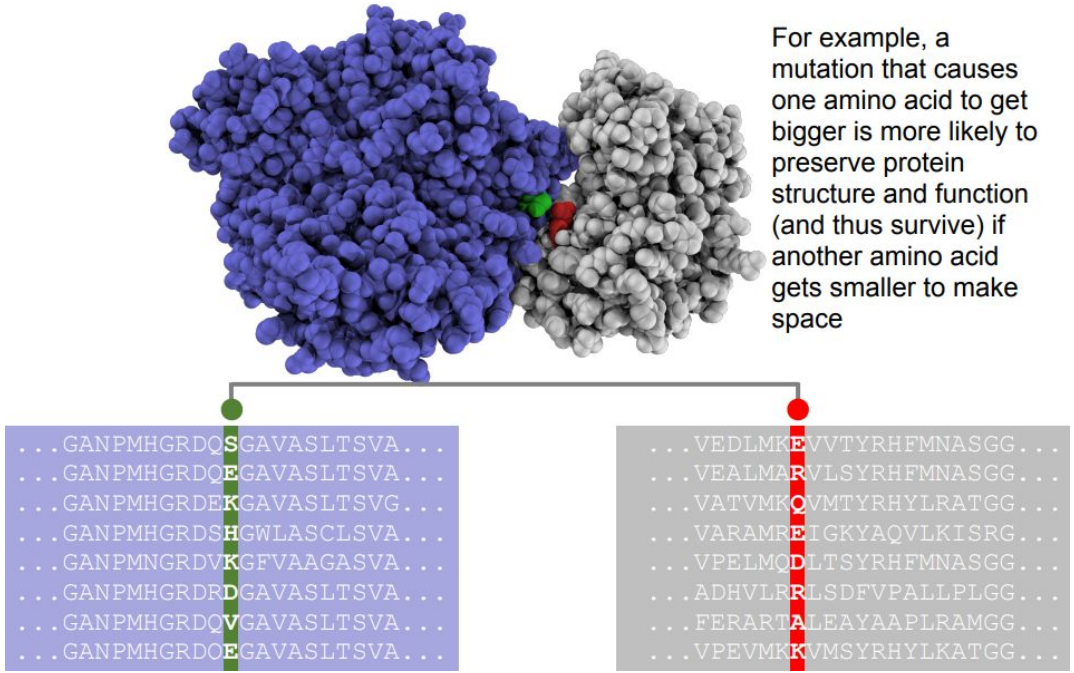


- Language models capture higher-order interactions *implicitly*

MSKGEE??TGVVPI????DGDVNGHKFSVY

PLMs are trained on natural sequence data …

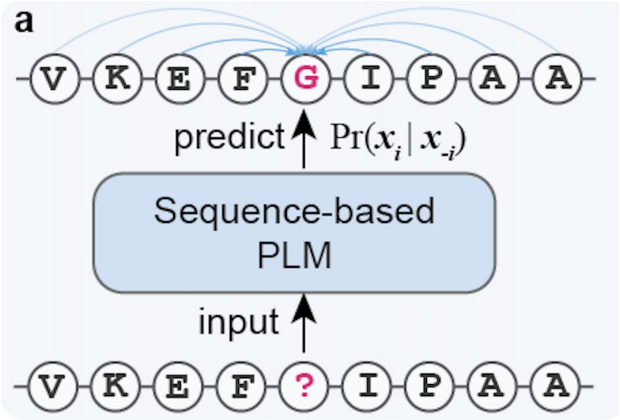
but we have seen that structure can have an “evolutionary effect” on protein sequences



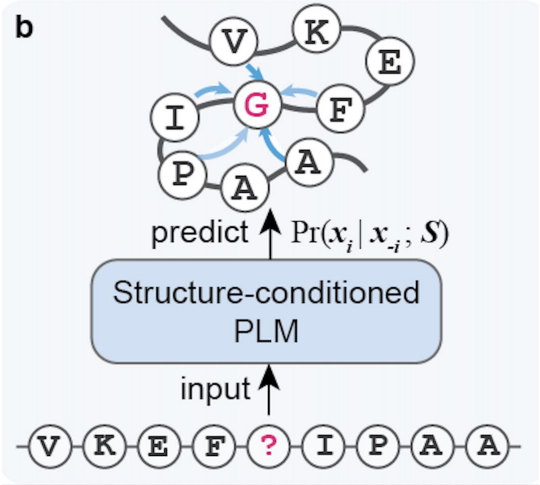
Can we leverage structure data in protein language models?

Structure-based PLM

Traditional PLM (sequence-based)



Structure-based PLM



Paper #1

RESEARCH

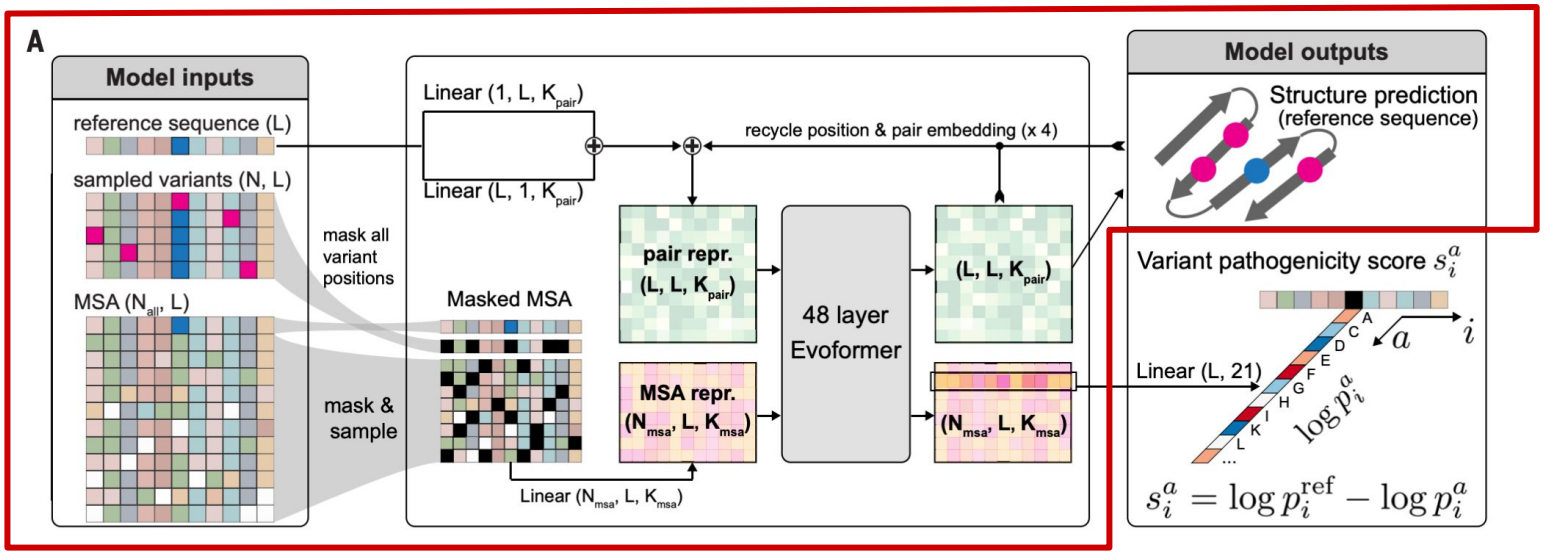
RESEARCH ARTICLE SUMMARY

MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

Jun Cheng*, Guido Novati, Joshua Pan†, Clare Bycroft†, Akvilė Žemgulytė†, Taylor Applebaum†, Alexander Pritzel, Lai Hong Wong, Michal Zielinski, Tobias Sargeant, Rosalia G. Schneider, Andrew W. Senior, John Jumper, Demis Hassabis, Pushmeet Kohli*, Žiga Avsec*

Almost identical to AlphaFold



Lecture:

Topic	Contents
Introduction	Introduction & Logistics
Basics in computational biology	Molecular biology
	No class (MLK day)
	Sequence alignment I
	Sequence alignment II
ML foundations	No Class (PyTorch video + exercise)
	Regression & Gradient descent
	Classification & Toolbox for Applied ML
	Neural networks
	Deep learning
Learning from sequence data	Deep learning for Protein/DNA sequences
	Large language models (LLMs)
Learning from high-dim data	Clustering and dimensionality reduction
	Generative AI
Learning from network data	Network basics & ML for graphs
	Graph neural network
Learning from structure data	Protein structure prediction & protein design
Advanced topics: ML for sequence data	Protein language models for prediction and generation
	Disease variant prediction
	Protein function prediction
	No class (Spring break)
	No class (Spring break)
Advanced topics: ML for structure data	Deep learning for structure prediction
	GNN for 3D structures
	Deep learning for structure generation
Advanced topics: ML for network data	Embeddings (representation learning)
	ML for protein design
	ML for drug discovery
Advanced topics: ML for high-D data	Dim reduction in bio data
	ML for system bio
	ML-guided biological discovery

Biology background

Intro to ML and DL

ML for bio data

ML in CompBio research

Hands-on exercise:

Announcements: Midterm Survey

- Informal Midterm Course Feedback (anonymous)
  - Provide mid-term feedback for this course
  - Suggest your favorite topics! May be incorporated in the remaining lectures
- Access the survey form by one of the following:
  - https://tinyurl.com/mlb-ief-s24
  - Canvas -> Syllabus -> “Midterm Survey”
  - QR Code: