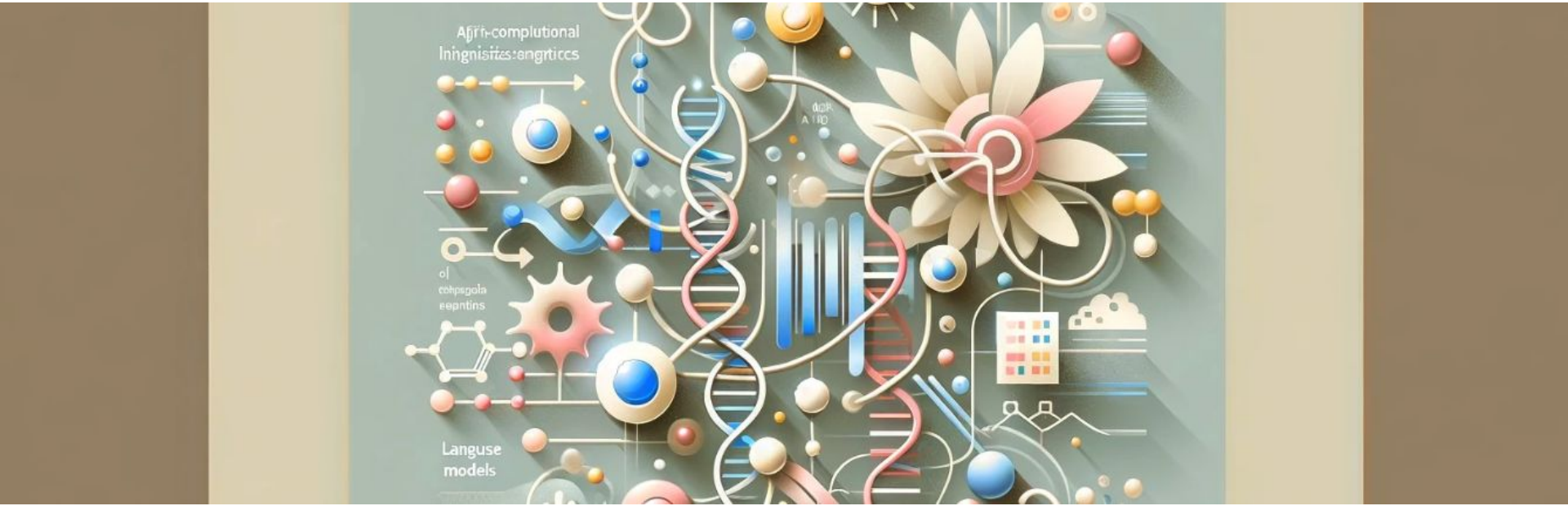
CSE7850/CX4803 Machine Learning in Computational Biology



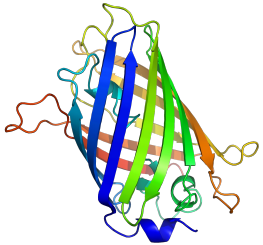
Lecture 16: Protein Language Models

Yunan Luo

Protein language models

Autoregressive language models

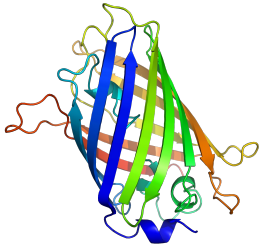
$$p(x) = \prod_{i=1}^{L} p(x_i | x_1 \dots x_{i-1})$$



MSKGEELFTGVVPILVELDGDVNGHKFSV_ ? S F Y V K

Masked language models

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



S F Y V K

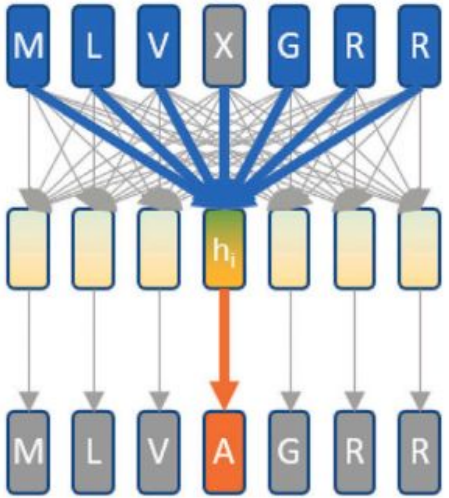
? MSKGEELFTGVVPI_VELDGDVNGHKFSVY

Can also mask multiple AAs at a time: MSKGEE??TGVVPI????DGDVNGHKFSVY

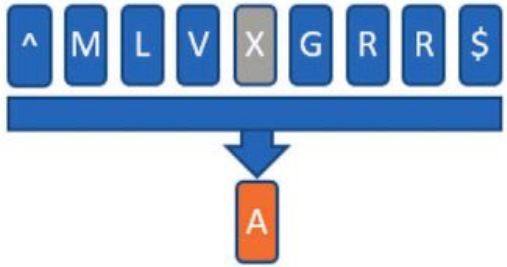
How to train a protein LM?

- Given a protein sequence $x = (x_1, x_2, \ldots, x_L)$
  - Mask a token x_i randomly (or multiple tokens)
  - Use other tokens as input
  - Train the neural network to predict x_i
  - Objective: $\mathbb{E}_{x\sim X}\mathbb{E}_M \sum_{i\in M} -\log p(x_i|x_{/M})$
- This training strategy is called **self-supervised learning**
  - Given only data (X), no labels (Y)
  - Simulate labels from the X itself
  - The model trains itself to learn one part of the input from another part of the input

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



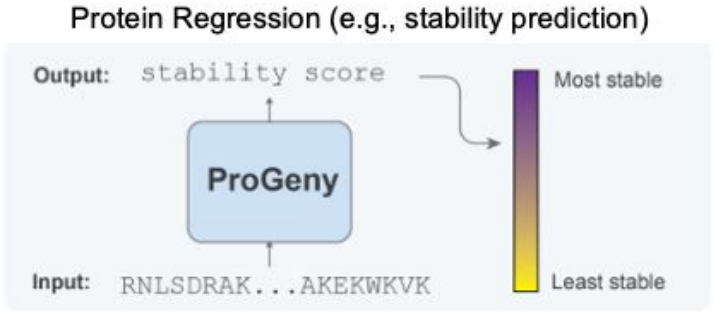
Processes whole sequence

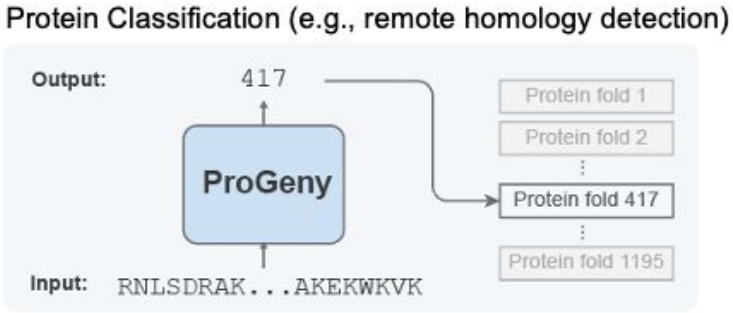


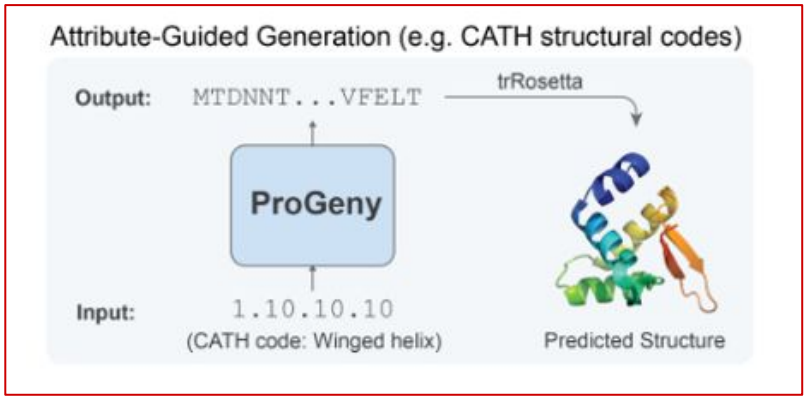
$$p(x_i = A|x_1 \ldots x_{i-1}, x_{i+1} \ldots x_L)$$

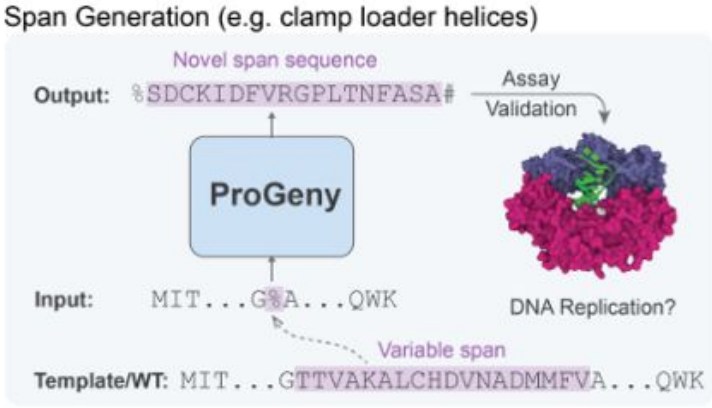
Bepler and Berger, *Cell Systems*, 2021

Application of protein language models

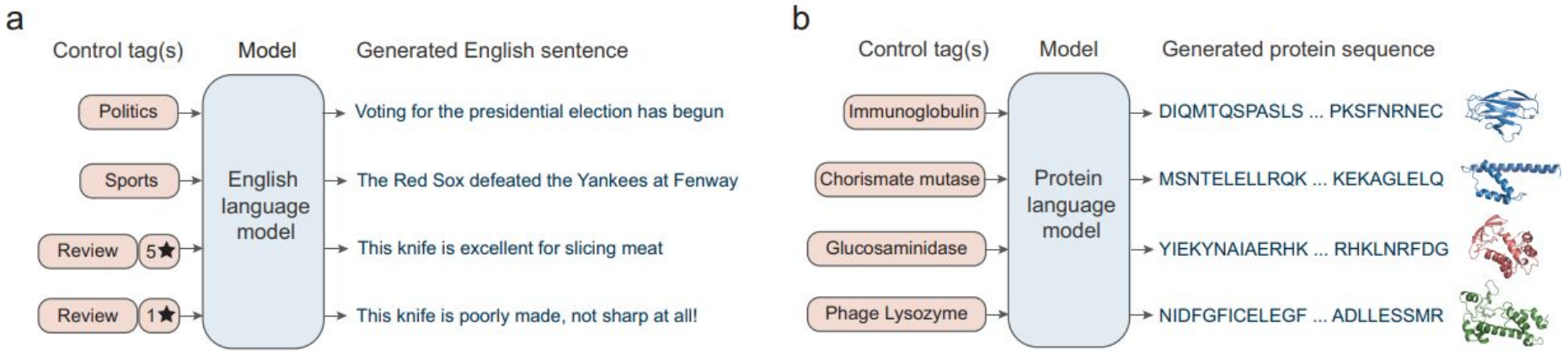






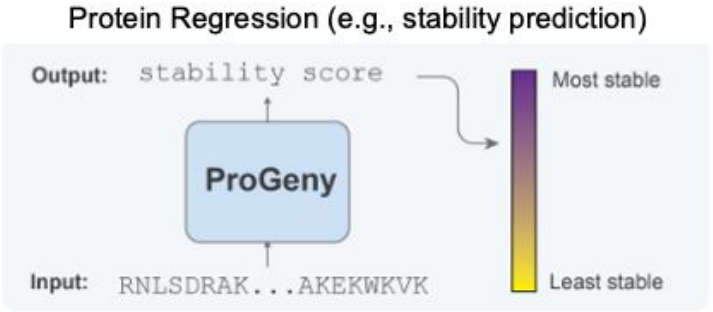


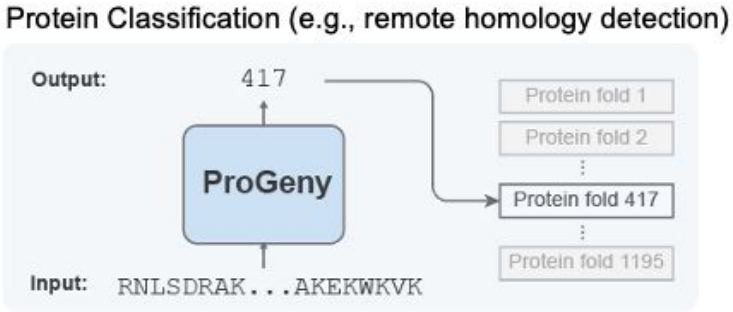
Application: generating novel functional protein sequences

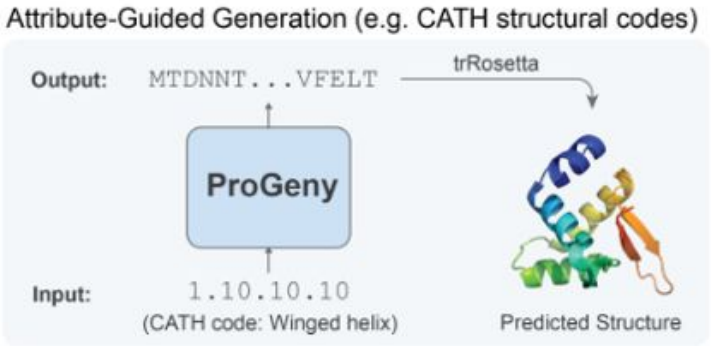


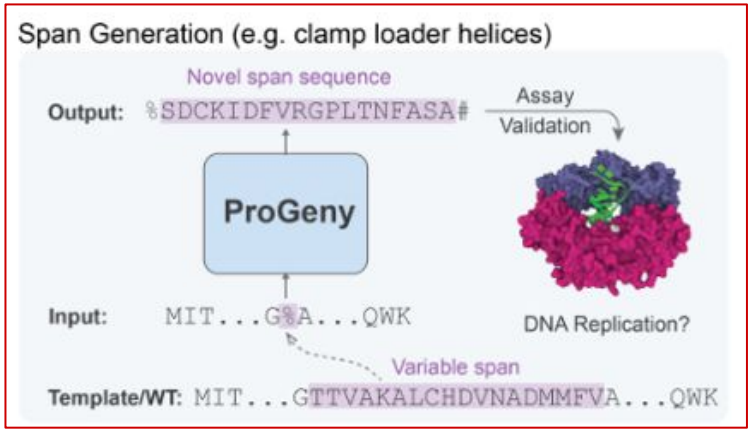
ProGen: Controllable generation of protein sequences
(Madani et al., Nature Biotech, 2023)

Application of protein language models

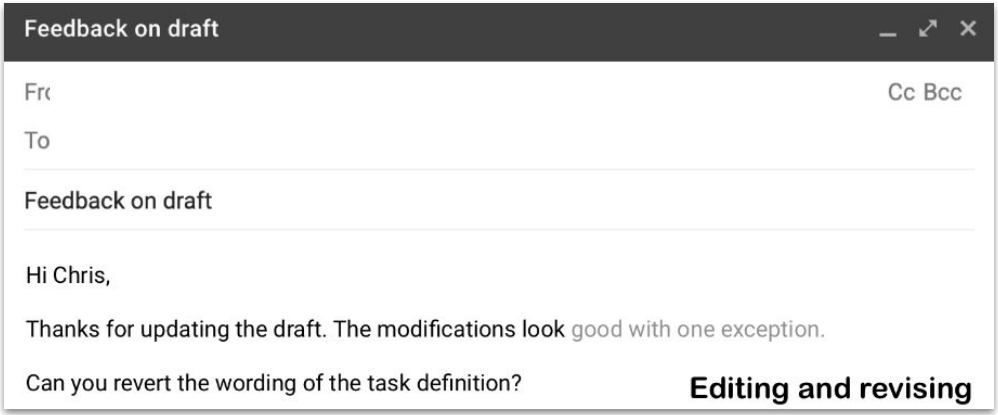


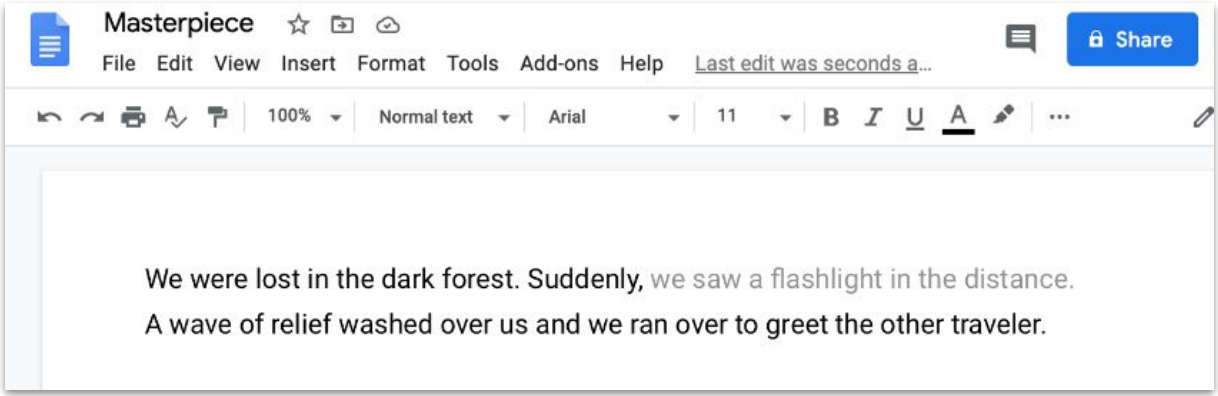






Fill in the Blanks using LMs





Fill in the Blanks using LMs

Fill in the blanks?

Consider the following sentence with blanks:

> *She ate ___ for ___*

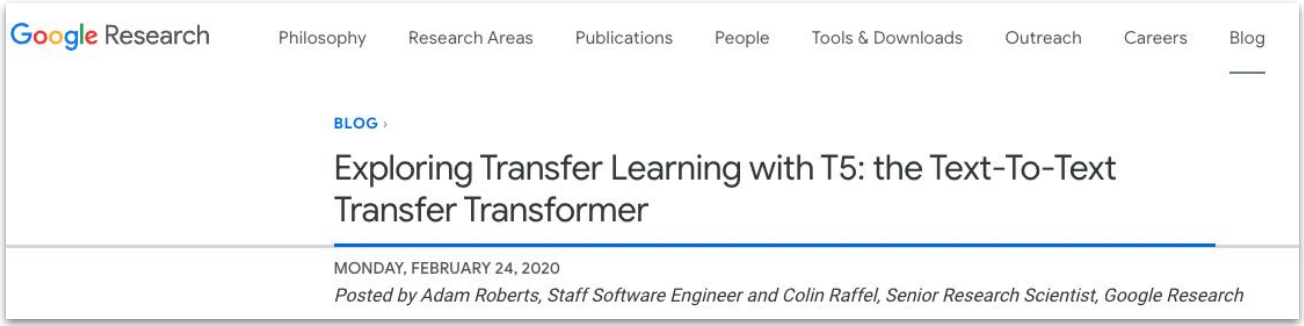
To fill in the blanks, one needs to consider both preceding and subsequent text (in this case, *"She ate"* and *"for"*). There can be many reasonable ways to fill in the blanks:

> *She ate **leftover pasta** for **lunch***
>
> *She ate **chocolate ice cream** for **dessert***
>
> *She ate **toast for breakfast before leaving** for **school***
>
> *She ate **rather quickly** for **she was in a hurry that evening***

The task of filling in the blanks is known as *text infilling* in the field of Natural Language Processing (NLP). It is the task of predicting blanks (or missing spans) of text at any position in text.

- This is exactly what LM does!
  - The model trains itself to learn one part of the input from another part of the input

Fill in the Blanks using LMs



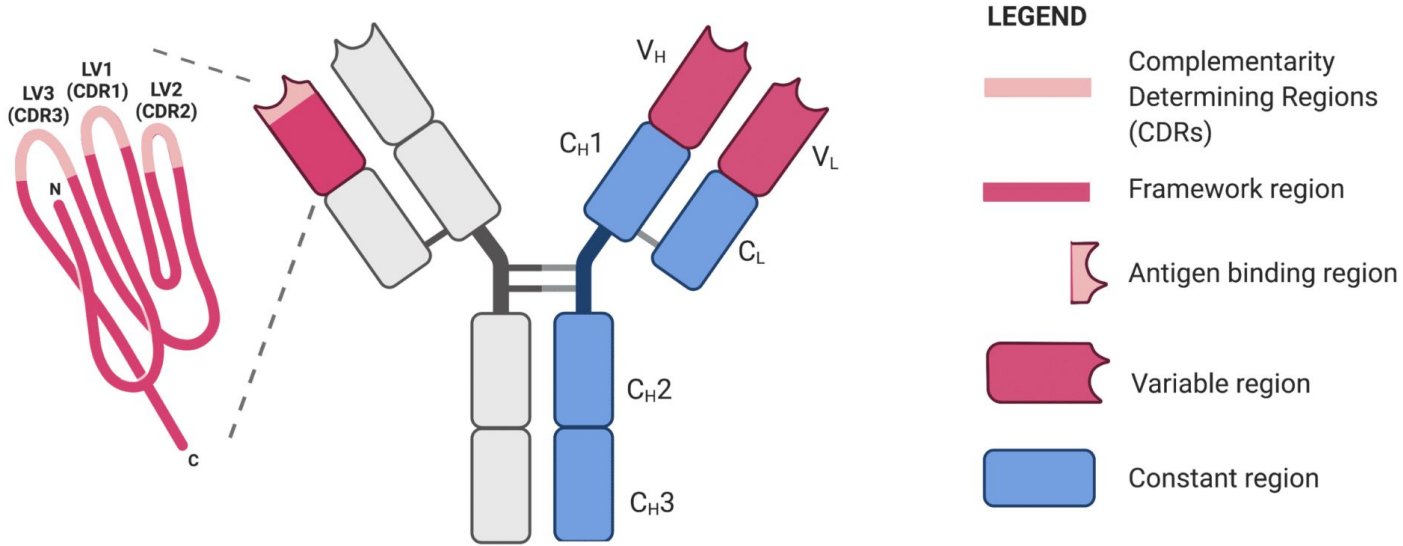
Input: I love peanut butter and __ sandwiches

N=1	N=2	N=4	N=8	N=16	N=32	N=64	N=512

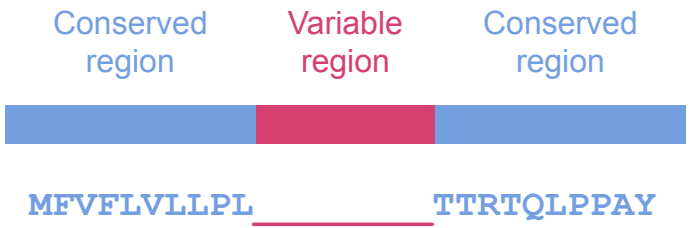
I love peanut butter and *jelly* sandwiches.

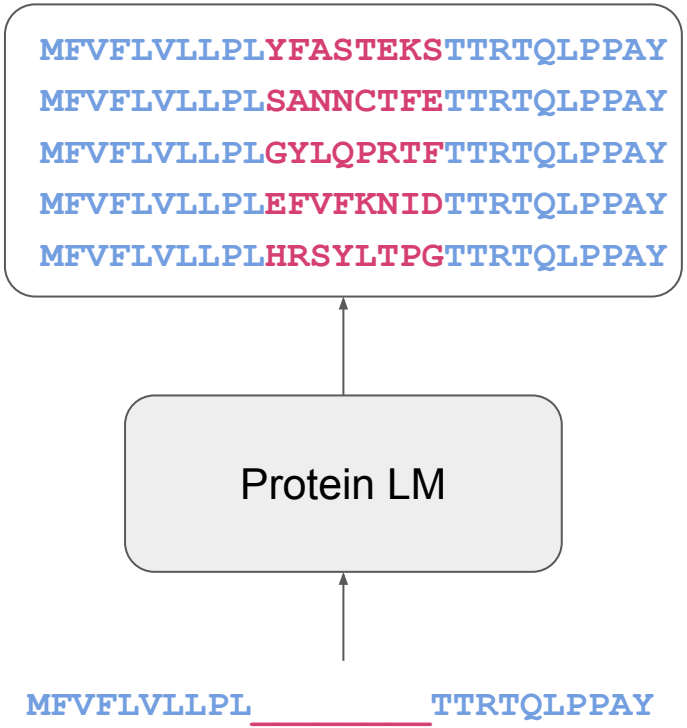
Fill in Blanks in Protein Sequences!

Application: antibody design



Application: Antibody Design

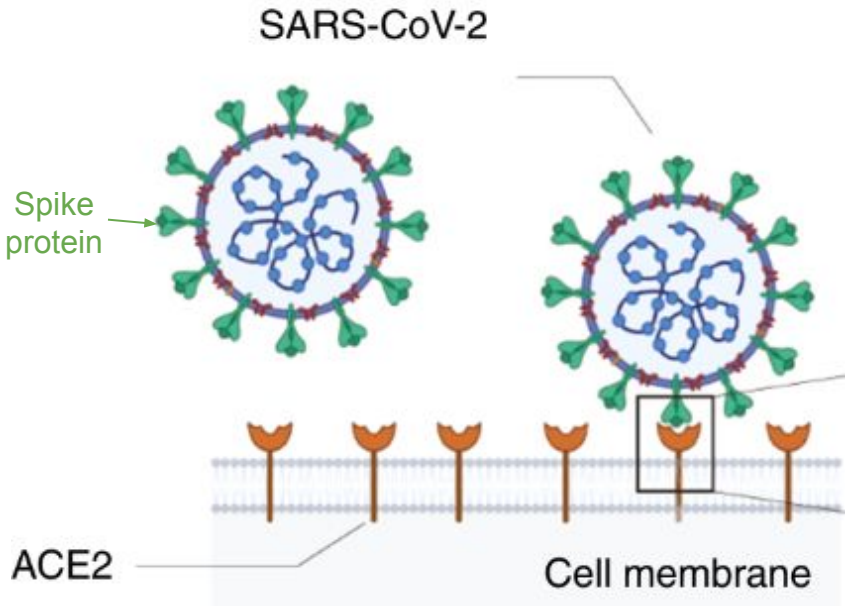




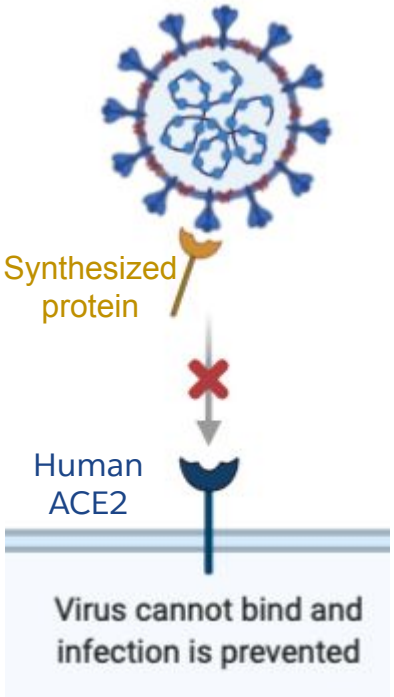
Sequences on this slides are not real antibody sequences

“Cure” COVID-19 using PLM

How SARS-CoV-2 gets into human cells?

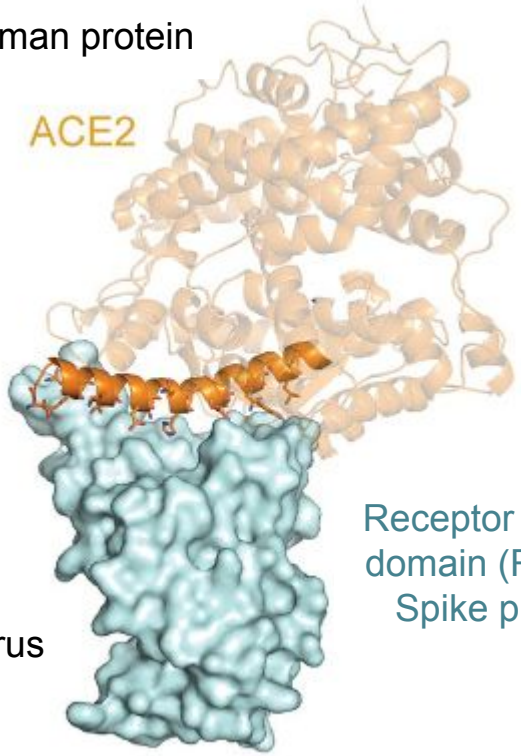


Blocking the virus using a “placeholder”



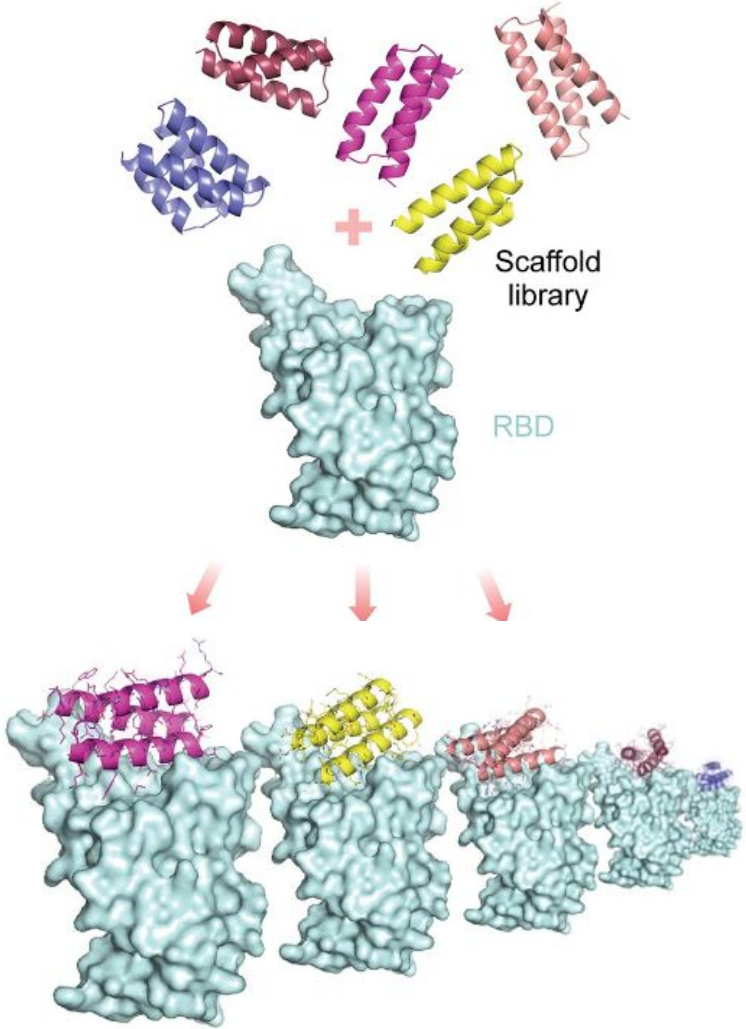
Design artificial binders

Human protein



Receptor binding domain (RBD) of Spike protein

Virus



Using PLM to generate binder sequences