MUTAGENIC: An Embedding-Based Approach to Protein Masking for Functional Redesign

Robin Pan*[1], Richard Zhu*[1], Vihan Lakshman[2], Fiona Qu[1]

[1]Harvard University, [2] Massachusetts Institute of Technology, *These authors contributed equally

Applying interpretability methods to protein design

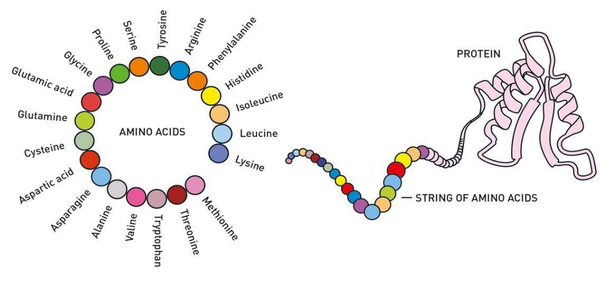


Image credits: https://www.nobelprize.org/uploads/2024/11/fig1_ke_en_24-5.pdf

Proteins are made up of 20 amino acids. Their specific combination and length (ie: primary sequence) dictate structure and function.

We draw inspiration from natural language processing counterfactuals, minimally modified inputs that alter the model's prediction. We aim to develop a model to generate a counterfactual equivalent for protein language models, minimally altered protein sequences with an altered function.

Original *(positive):*
That movie was so **exciting**

Counterfactual *(negative):*
That movie was so **boring**

Original *(function A)*:
LTQSPSSLAVSAG**ERVTM**....

Counterfactual (function B):
LTQSPSSLAVSAG**KLLAR**....

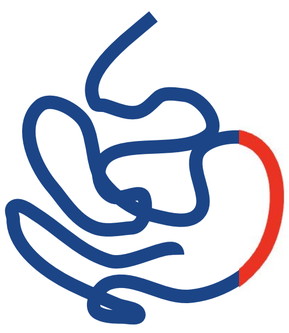
The framework

Masking Model
Selects which amino acids to mask

→

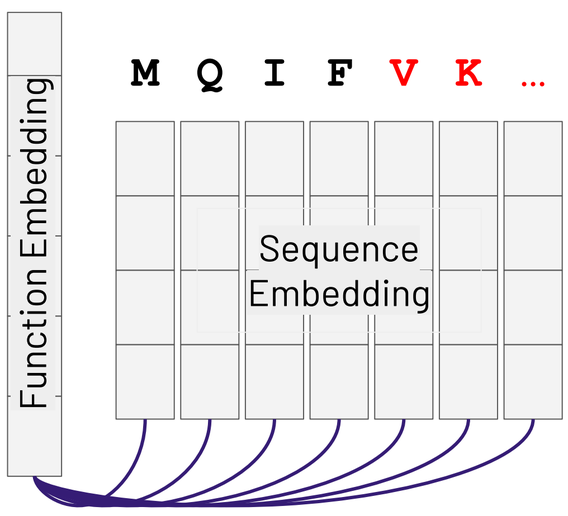
Editor Model
Fills in masked tokens with protein residues (ESM3)

Our masking model



MQIFVKTLTGK...
Targeted Function:
IPR020568

→ ESM3 →



Cosine similarities

↓

← ESM3 ←

M Q I F <mask><mask> ...

0.8	0.7	0.5	0.9	0.0	0.1

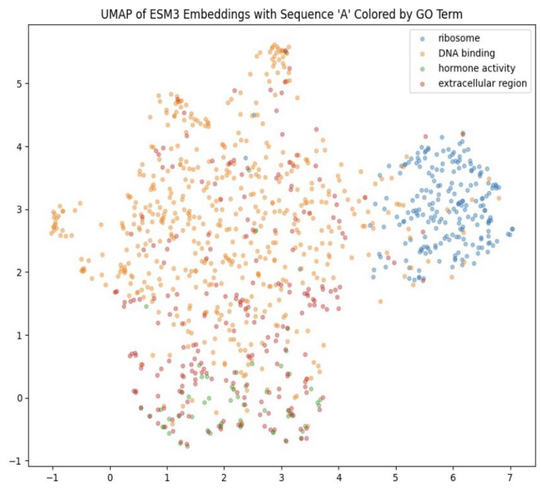
MQIFSTLHL...
IPR020568

How do we choose what sites to mask?

1. Embed the InterPro term corresponding to target function
2. Generate protein embedding
3. Mask sites with the lowest cosine similarity to the target embedding

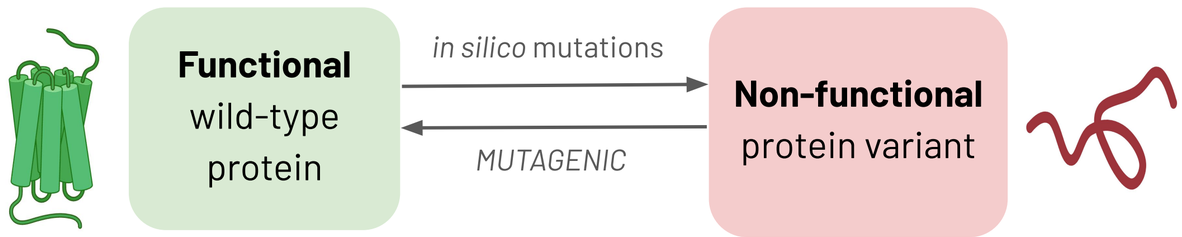
ESM3 embeddings can cluster different functions

Question - Can we use ESM3 functional embeddings in our masking model?
Experiment - Generate functional embeddings of distinct function Interpro terms with ESM3
Result - Similar function terms cluster together



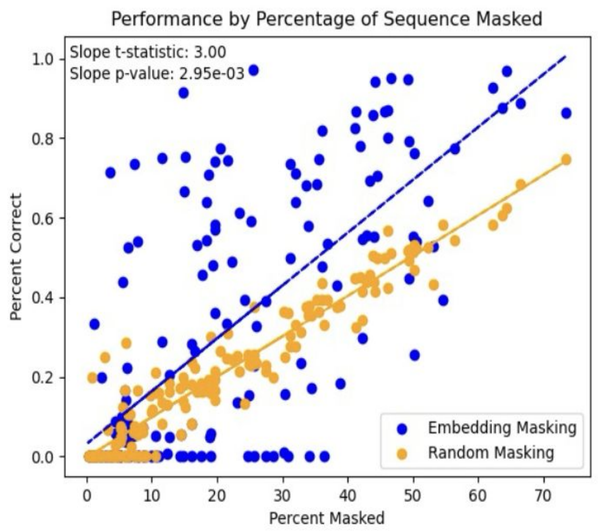
In-silico datasets were used for validation

For evaluation of change of function - we focused on **gain of function** mutations of 200 randomly sampled proteins with artificial substitutions.



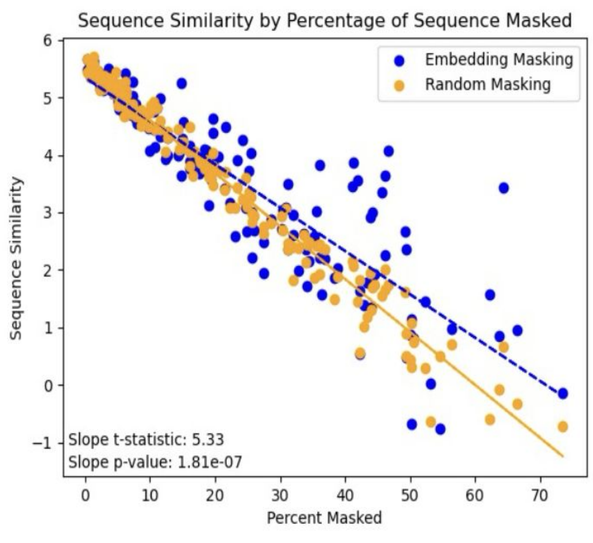
Why? Novel and distinct functional changes of proteins still not a routine engineering objective in the field.

Results



Sequence similarity between edited sequence and functional wild-type sequence quantified is measured via BLOSUM80 matrix (higher score is better).

Our model identifies chunked mask deletion sites better than a **random masker**.



Future directions of work

- Number of masked token: We will dynamically select the optimal masking percentage for each protein based on a threshold for similarity (or change in similarity) with target embedding
- Scoring model: Adding a scoring model that can continually communicate with and learn the number of tokens to mask and where

References:

ESM3: Hayes, Thomas, et al. *Simulating 500 Million Years of Evolution with a Language Model*, 2 July 2024, https://doi.org/10.1101/2024.07.01.600583.
Wang, Yongjie, et al. "A survey on natural language counterfactual generation." Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 4798–4818, https://doi.org/10.18653/v1/2024.findings-emnlp.276.