MUTAGENIC: AN EMBEDDING-BASED APPROACH TO PROTEIN MASKING FOR FUNCTIONAL REDESIGN

Robin Pan* **& Richard Zhu***
Harvard University
{rpan,rzhu}@college.harvard.edu

Vihan Lakshman
MIT CSAIL
vihan@mit.edu

Fiona Qu
Department of Systems Biology
Harvard University
fionaqu@g.harvard.edu

1 INTRODUCTION

Recent advances in language models have been applied to protein sequences because of their critical functions in biological processes and the availability of large datasets. Protein engineering has already proven to be impactful in areas such as therapeutics, agriculture, the environment, and biomanufacturing (Brange et al., 1991; Vojcic et al., 2015). Motivated by the challenge of protein design, this paper investigates the following question: *How can we efficiently identify residues to edit in the engineering of proteins with specific target functions*?

In this paper, we propose a novel embedding-based masking approach to edit a given protein to achieve a new target function. More formally, let $F = \{f_1, f_2, \ldots, f_n\}$ denote the set of possible protein functions. Given a protein sequence $s = s_1 s_2 \ldots s_N$ composed of amino acids $\{s_i\}_{i=1}^N$ with function $f \in F$ and a target function $f' \in F$, our goal is to return a new protein sequence s' with functionality f'.

Existing masking approaches for protein design rely on the user to select masking sites or are specific to the use case. Vincoff et al. (2024) design fusion oncoproteins by masking residues with high probability of participating in protein-protein interactions as predicted by SaLT&PepPr (Brixi et al., 2023). Various masking strategies have been developed for interpretability and feature attribution in human language models (Ross et al., 2020; Barkan et al., 2024) and, to a lesser extent, biological sequence models (Linder et al., 2022). To our knowledge, our framework MUTAGENIC presents the first use of vector embeddings to determine masking sites in protein language models, and the only masking model with capabilities for generalized function-guided design.

2 METHODOLOGY AND RESULTS

We present MUTAGENIC[1], a framework for generalized protein function modification. In this pipeline, we utilize ESM3 (Hayes et al., 2024), a foundation model that represents proteins using 6 tracks, including sequence, structure, and function. By providing a sequence with masked tokens, ESM3 is able to unmask each token along any of these "tracks" to fully generate the protein. In MUTAGENIC, we start with a protein sequence of N amino acid residues: $\{r_1, ..., r_N\}$. We identify the optimal residues to mutate conditioned upon the protein's original function f and our target function f' (with the functions specified as labels from the InterPro database (Hunter et al., 2009)).

Specifically, we use the ESM3 encoder to create a vector embedding of each amino acid in the protein sequence $\{\mathbf{v}_{r_1}, ..., \mathbf{v}_{r_N}\}$, as well as embeddings of the original and target functions, $\mathbf{v}_f$ and $\mathbf{v}_{f'}$. The vectors $\{\mathbf{v}_{r_1}, ..., \mathbf{v}_{r_N}, \mathbf{v}_f, \mathbf{v}_{f'}\}$ are all generated using ESM3's function "track", so we call them *functional embeddings*. Next, we calculate s_{if}, the similarity between residue i's functional embedding and the original function f's embedding through cosine similarity: $s_{if} \propto \mathbf{v}_{r_i} \cdot \mathbf{v}_f, \forall i$. The value $s_{if'}$ is defined analogously to capture the per-residue similarity with f'. Now, we combine these similarities into scores: $\text{score}_i = g(s_{if}, s_{if'})$, where g is a general aggregation function tailored to each protein engineering task. We then select the k residues with the highest score, where n is a user-specified number of mutation sites. These k sites are masked in the original protein sequence to yield a masked sequence. ESM3 then fills in these masked sites with new residues conditioned upon the target function f'. This yields an edited protein sequence more closely aligned with f'. Figure 1 shows the full pipeline.

*These authors contributed equally to this work.

[1] https://github.com/vihan-lakshman/mutagenic-experiments

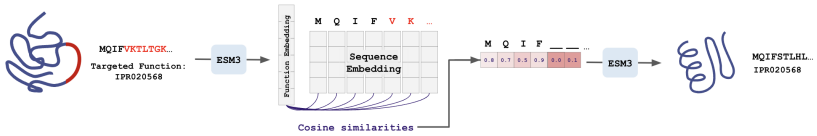

Figure 1: The MUTAGENIC pipeline combines ESM3 with embedding-based masking to mutate residues that will align the protein sequence more closely with a target function f'.

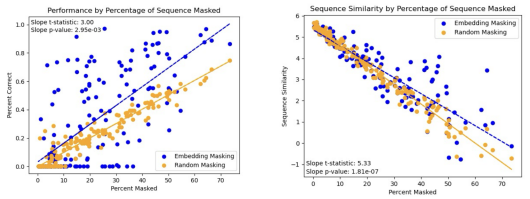

Figure 2: MUTAGENIC performs better than the baseline in mask site selection (left) and sequence similarity of reconstructed sequence (right).

To evaluate MUTAGENIC, we first verify that ESM3's functional embeddings capture meaningful biological relationships, a core assumption in our pipeline. We performed a UMAP (McInnes et al., 2018) on the ESM3 functional embeddings for all InterPro labels associated with four general biological functions: "ribosome", "DNA binding", "hormone activity", and "extracellular region". The UMAP results (Figure A in Appendix) showed strong separation between the embeddings for distinct functions spatially separated in the cell (ribosome, DNA binding, and extracellular region). However, there was heavy overlap between embeddings for hormone activity and extracellular region, which aligns with the fact that many hormones travel extracellularly in the body. This suggests ESM3 functional embeddings are good representations of protein function.

Next, we demonstrate that MUTAGENIC can reverse massive substitutions and restore wildtype protein function better than a random masking baseline. We generated an *in silico* dataset of 200 functional proteins with large contiguous regions of random substitution mutations, rendering proteins non-functional. The number of residues altered for a given protein P varied from 0 to $> 70\%$ of the sequence length. Using MMseqs2 sequence clustering (Steinegger & Söding, 2017), we ensured the dataset contains enormous functional diversity. To apply MUTAGENIC to this problem, we define f' as the InterPro labels for the wildtype function, while f is not defined since the mutated proteins are assumed to be non-functional. Then, the score is $\text{score}_i = -s_{if'}$ for each residue i, as we select residues most dissimilar to the target function f' for mutation. For each protein P, the n_P residues with the highest scores were masked. We compare against a random masking baseline, where n_P residues are randomly selected throughout the protein sequence for mutation. We choose random masking due to a lack of other masking models that can identify optimal mutation sites across a general range of functions—a core strength of our pipeline. Finally, we use reconstruction of the wildtype sequence as a proxy for the degree to which MUTAGENIC and the baseline method recover wildtype function. To do this, we use BLOSUM80 substitution matrices (Henikoff & Henikoff, 1992) to capture the similarity between the wildtype sequence and output sequence.

In Figure 2, we show that the embedding-based masking in MUTAGENIC (blue) masks a higher percentage of mutated residues in our test set compared to random masking. Moreover, the difference in accuracy increases with the percent of the wildtype protein sequence mutated (x-axis), implying that MUTAGENIC is able to identify mutated residues dissimilar to the wildtype function better (compared to baseline) as the proportion of mutated residues increases. Using a 2-sample *t*-test, we find that this rate of increase in performance of MUTAGENIC (slope) is significantly different than the rate of increase for the baseline ($p < 0.01$), implying that MUTAGENIC's advantage over baseline increases as the proportion of wildtype protein mutated increases. We also show in Fig. 2 that the wildtype sequence recovery (as measured by BLOSUM80) of the final ESM3-edited protein sequence is higher when using MUTAGENIC's embedding-based masked sites compared to randomly masked sites. Once again, the difference in slope is statistically significant, implying that MUTAGENIC's ability to recover wildtype function increases with the percent of protein mutated.

MEANINGFULNESS STATEMENT

Protein language models are based on the concept that underlying patterns of protein evolution can be captured through their sequences alone. Our masking model, in particular, leverages functional embeddings—numerical representations of protein function across a continuous space. We demonstrate that these embeddings can effectively differentiate functions within this space, and can be used to identify key residues for redesign. By leveraging these representations of life, our goal is to not just to understand, but to actively engineer the powerhouses of life.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 2313998. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. National Science Foundation.

REFERENCES

Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. Llm explainability via attributive masking learning. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9522–9537, 2024.

Jens Brange, Guy G Dodson, and Bing Xiao. Designing insulin for diabetes therapy by protein engineering. *Curr. Opin. Struct. Biol.*, 1(6):934–940, December 1991.

Garyk Brixi, Tianzheng Ye, Lauren Hong, Tian Wang, Connor Monticello, Natalia Lopez-Barbosa, Sophia Vincoff, Vivian Yudistyra, Lin Zhao, Elena Haarer, et al. Salt&peppr is an interface-predicting language model for designing peptide-guided protein degraders. *Communications Biology*, 6(1):1081, 2023.

Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, pp. 2024–07, 2024.

Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, et al. Interpro: the integrative protein signature database. *Nucleic acids research*, 37(suppl_1):D211–D215, 2009.

Johannes Linder, Alyssa La Fleur, Zibo Chen, Ajasja Ljubetič, David Baker, Sreeram Kannan, and Georg Seelig. Interpreting neural networks for biological sequences by learning stochastic masks. *Nature machine intelligence*, 4(1):41–54, 2022.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Alexis Ross, Ana Marasović, and Matthew E Peters. Explaining NLP models via minimal contrastive editing (MiCE). *arXiv [cs.CL]*, December 2020.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. ISSN 1087-0156. doi: 10.1038/nbt.3988. URL https://www.nature.com/articles/nbt.3988.

Sophia Vincoff, Shrey Goel, Kseniia Kholina, Rishab Pulugurta, Pranay Vure, and Pranam Chatterjee. Fuson-plm: A fusion oncoprotein-specific language model via focused probabilistic masking. *bioRxiv*, 2024. doi: 10.1101/2024.06.03.597245. URL https://www.biorxiv.org/content/early/2024/06/04/2024.06.03.597245.

Ljubica Vojcic, Christian Pitzler, Georgette Körfer, Felix Jakob, Ronny Martinez, Karl-Heinz Maurer, and Ulrich Schwaneberg. Advances in protease engineering for laundry detergents. *N. Biotechnol.*, 32(6):629–634, December 2015.

A APPENDIX

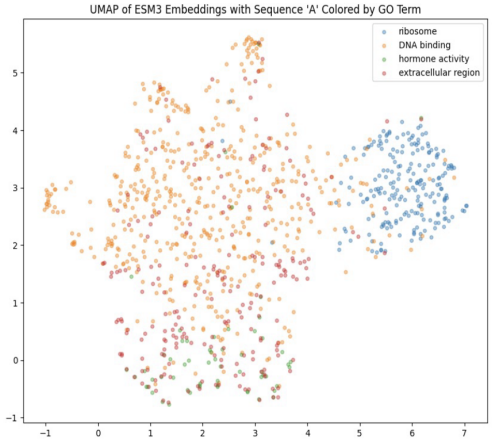

Figure 3: A UMAP of selected Gene Ontology function terms demonstrate that ESM3 functional embeddings can successfully distinguish between and relate different biological functions