Unsupervised language models for disease variant prediction

Allan Zhou*, **Nick Landolfi***, Dan O'Neill
Stanford University

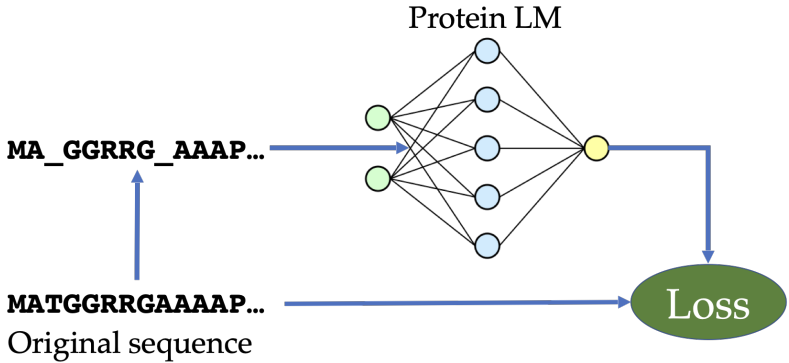
Revision: 4 December 2022, for Neurips Workshops; *Equal contribution

Pathogenicity prediction for protein variants of human genes

- **goal**: predict pathogenicity
- **challenge**: lack high-quality labels, infeasible to collect
- **approach**: use likelihood as a proxy for pathogenicity
  - *evolutionary principle*: *less frequently* occurring variants are *more likely pathogenic*
  - prior work: train generative models on multiple sequence alignments (MSAs)[1]
- **our work**: *pretrained language models* (LMs) predict pathogenicity comparably with state-of-the-art
  - *zero shot, no fine-tuning, no MSAs*
  - opens the possibility of flexibly scoring any variant

[1] Frazer et al., 2022 *Disease variant prediction with deep generative models of evolutionary data*

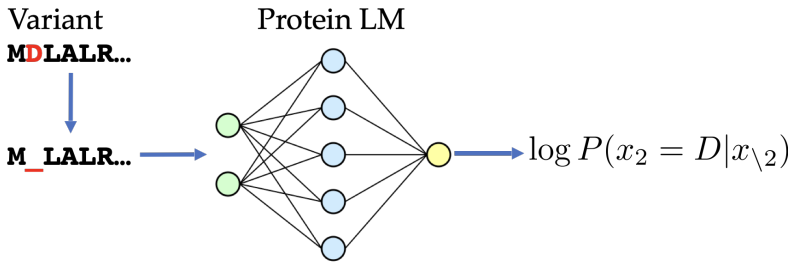
Training language models via self-supervision on large protein datasets



- typical approach: on natural sequences, train to predict randomly masked residues[2]
  - example dataset: UniRef50, consisting of 45 million protein sequences

[2] Elnaggar et al., 2021 *ProtTrans: towards cracking the language of life's code through self-supervised deep learning...*

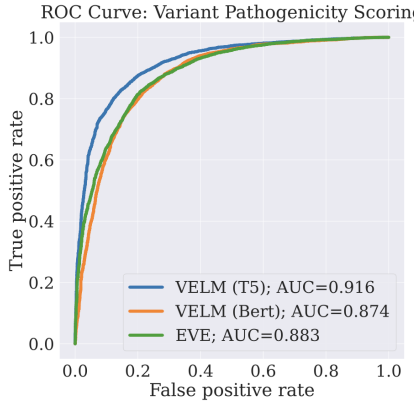
Compute conditional likelihood of mutated sequence

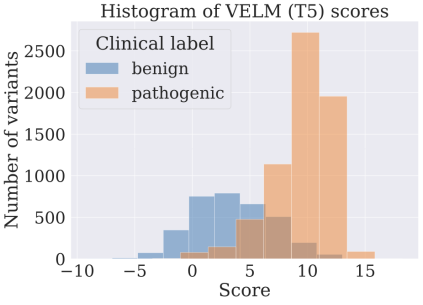


- define a *score* $S(x^{\text{mt}}) := \sum_{i \in M} \log P(x_i = x_i^{\text{wt}} \mid x_{\setminus M}^{\text{wt}}) - \log P(x_i = x_i^{\text{mt}} \mid x_{\setminus M}^{\text{mt}})$
  - here $M = \{i : x_i^{\text{mt}} \neq x_i^{\text{wt}}\}$ is the set of *mutated indices*
  - prior work uses this score for protein *function*[3]

[3] Meier et al., 2021; *Language models enable zero-shot prediction of the effects of mutations on protein function*

Zero-shot language models have better aggregate performance





- evaluate on *high-quality clinical labels* (ClinVar labeled variants with at least one star)
  - compare language models T5, Bert (Elnaggar et al., 2021) with state-of-the-art EVE (Frazer, 2021)
  - VELM with T5 has highest aggregate AUC, despite not using MSAs

Protein language models predict pathogenicity zero-shot

thank you! — please feel free to reach out to us at poster session or via email

1. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.
2. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al. ProtTrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
3. J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.
4. N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. Genome-wide prediction of disease variants with a deep protein language model. *bioRxiv*[4]

[4] Concurrent work, not referenced in talk.