Unsupervised language models for disease variant prediction

Allan Zhou, Nicholas C. Landolfi*, Daniel C. O'Neill*

Machine Learning for Structural Biology Workshop, NeurIPS 2022

Introduction

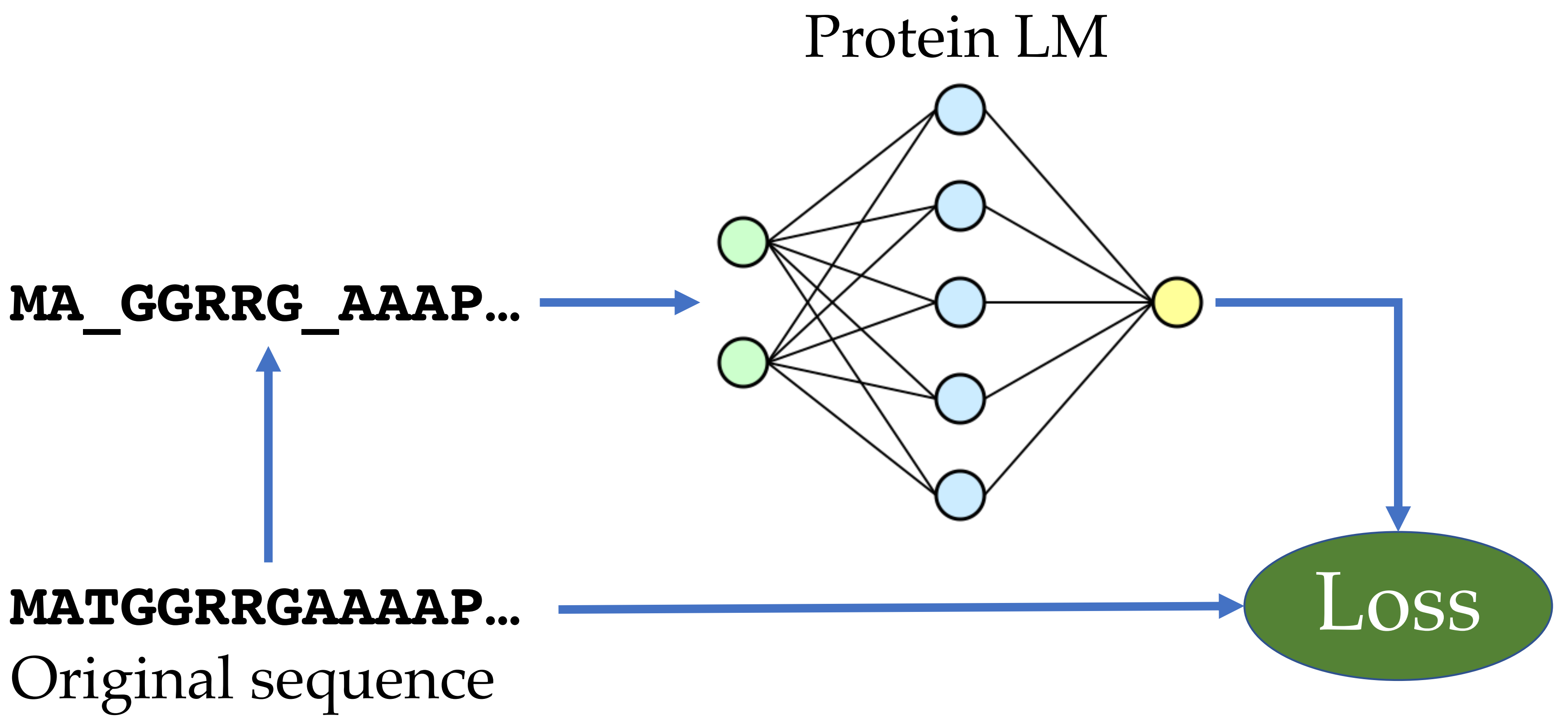
Predicting pathogenicity for protein variants in human genes suffers from a lack of high quality supervision (labels). *Unsupervised* methods predict protein sequence likelihood as a proxy for fitness (**evolutionary principle**). However, this typically requires training generative models on MSAs for each gene (Frazer et al, 2021).

Findings: Pretrained *protein language models* can score gene variant pathogenicity zero-shot, without data preprocessing or finetuning on per-gene MSAs. We call this unsupervised protein LM scoring method **VELM** and show that it performs comparably to state of the art methods on clinically labeled gene variants.

Background: Protein Language Models (LMs)

Protein LMs are trained by self-supervised learning on large open datasets of protein sequences.

A typical training objective is for the LM to predict the missing amino acids in a randomly masked sequence. This makes them ideal for predicting variant likelihood, as a proxy for evolutionary fitness.



Protein LM

MA_GGRRG_AAAP...

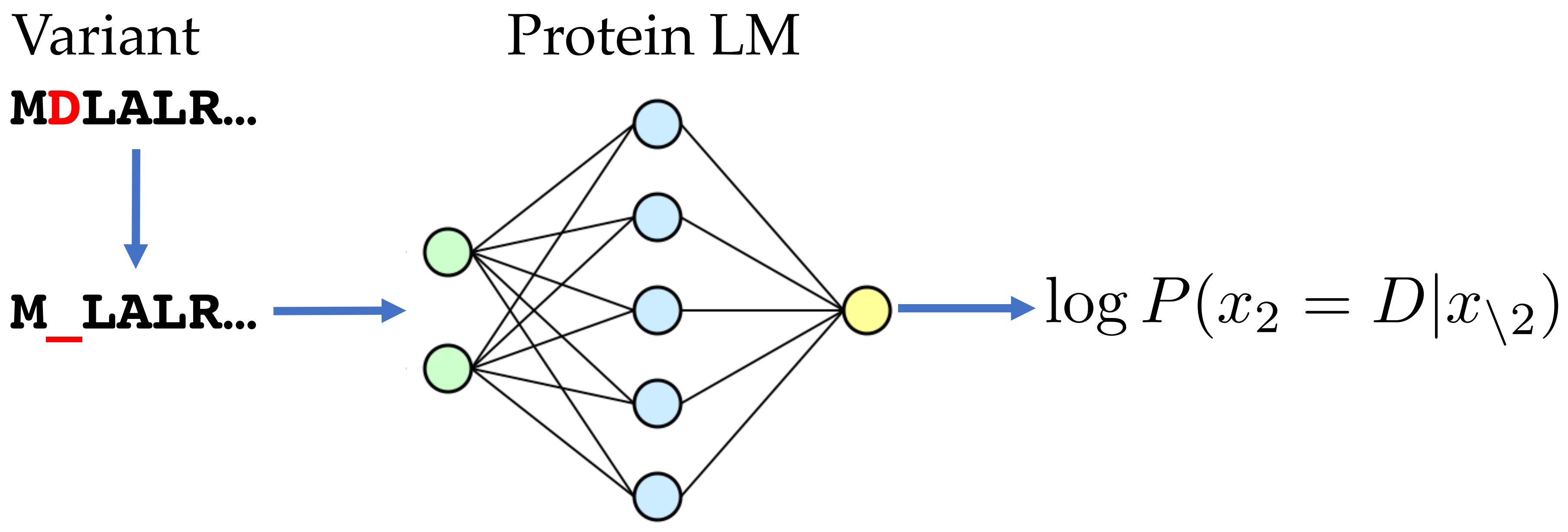
MATGGRRGAAAAP...

Original sequence

Loss

Method

To score a missense variant, we mask the sequence at the mutated location and output the protein LM's conditional probability distribution at that location:



Variant

MDLALR...

M_LALR...

Protein LM

$\log P(x_2 = D | x_{\setminus 2})$

To define a pathogenicity score, we evaluate the log odds ratio between variant and wildtype at the mutated positions (Meier et al, 2021).

$$S(x^{\text{mt}}) := \sum_{i \in M} \log P(x_i = x_i^{\text{wt}} | x^{\text{wt}}_{\setminus M}) - \log P(x_i = x_i^{\text{mt}} | x^{\text{mt}}_{\setminus M})$$

$$M = \{i : x_i^{\text{mt}} \neq x_i^{\text{wt}}\}$$

Intuitively, S( ) is *higher* when the mutations make the variant sequence *less likely* than the wildtype, making it more likely to be pathogenic.

References

[1] 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 lifes code through self-supervised deep learning and high performance computing. IEEE transactions on pattern analysis and machine intelligence, 2021.

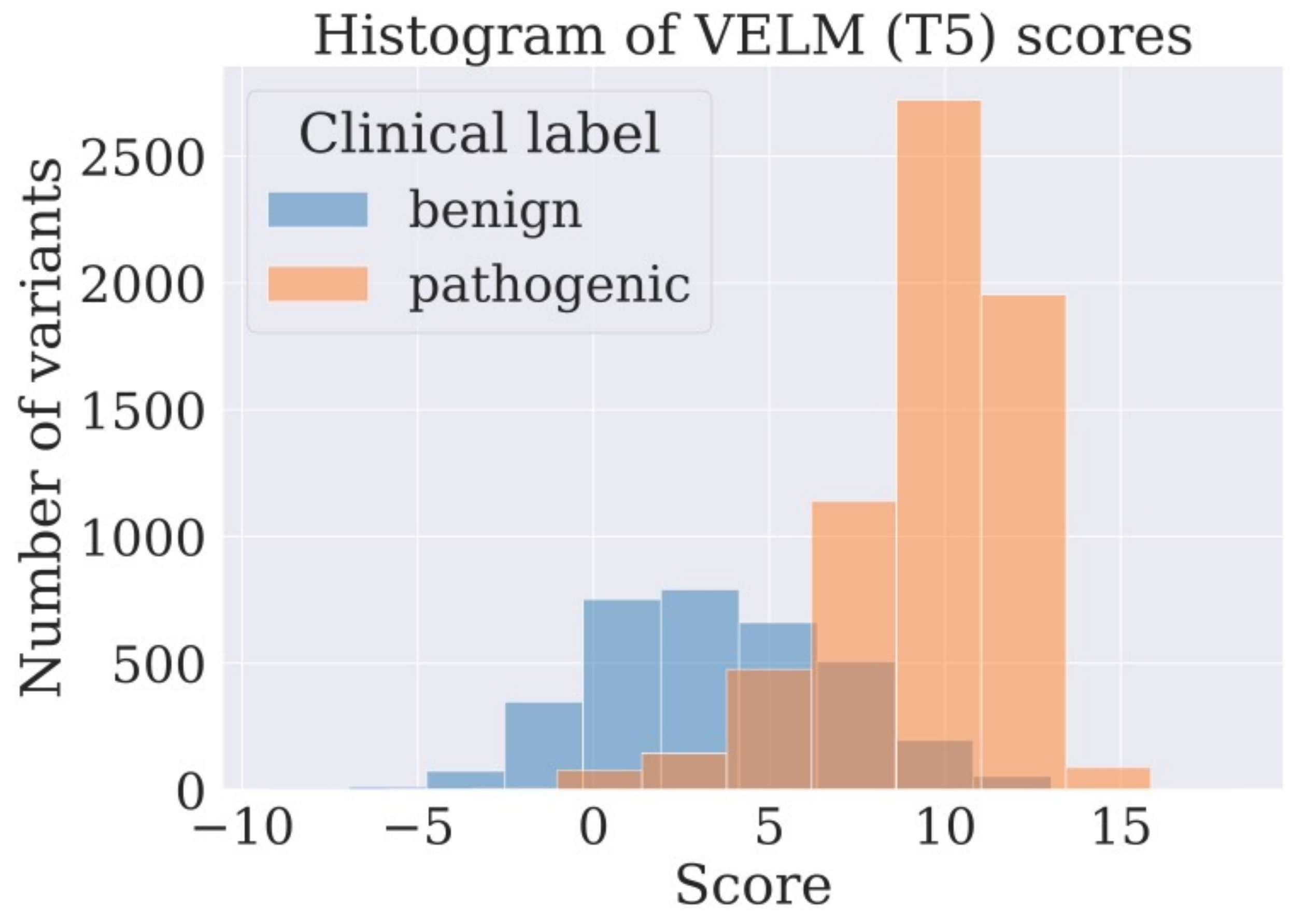
[2] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. Nature, 599(7883):91–95, 2021.

[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.

Experiments

Using Prot-T5 (Elnaggar et al, 2021), we evaluate VELM on a set of protein variants with known clinical labels.

VELM's pathogenicity score largely separates the variants by clinical label (benign vs. pathogenic), without requiring finetuning on gene-specific data.



VELM performs comparably to EVE (Frazer et al, 2021), which trains a separate generative model per-gene. The performance is closest on genes with more clinical labels (less noise).

Metric	VELM (T5)	EVE
mAUC (≥ 1 labels)	0.901	0.917
mAUC (≥ 3 labels)	0.912	0.930
mAUC (≥ 5 labels)	0.933	0.936